Signatures of the neurocognitive basis of culture wars found in moral psychology data

Nestor Caticha and Renato Vicente

March 11, 2010

Moral Foundation Theory (MFT)\(^1\) states that groups of different observers may rely on partially dissimilar sets of moral foundations, thereby reaching different moral valuations on a subset of issues. With the introduction of functional imaging techniques, a wealth of new data on neurocognitive processes has rapidly mounted and it has become increasingly more evident that this type of data should provide an adequate basis for modeling social systems\(^2\). In particular, it has been shown that there is a spectrum of cognitive styles with respect to the differential handling of novel or corroborating information. Furthermore this spectrum is correlated to political affiliation\(^3\). Here we use methods of statistical mechanics\(^4\) to characterize the collective behavior of an agent-based model society whose interindividual interactions due to information exchange in the form of opinions, are in qualitative agreement with neurocognitive and psychological data\(^5\)-\(^13\). The main conclusion derived from the model is that the existence of diversity in the cognitive strategies yields different statistics for the sets of moral foundations and that these arise from the cognitive interactions of the agents. Thus a simple interacting agent model, whose interactions are in accord with empirical data about moral dynamics, presents statistical signatures consistent with those that characterize opinions of conservatives and liberals. The higher the difference in the treatment of novel and corroborating information the more agents correlate to liberals.

The proponents of MFT have identified at least five moral foundations or dimensions that, potentially, are universally present in humans. These dimensions are manifested in different manners, not only across time and cultures, but also within a society. Individuals with different attributions of the relative importance of the dimensions will be led to fundamental misunderstandings of moral motivations of each other. Within the MFT, extensive empirical support\(^14\)-\(^15\) has been gathered for the fact that the use of different subsets of moral foundations by groups is significantly correlated with a coordinate characterizing the group along the political spectrum. The subsets are such that liberals tend to rely more strongly on aspects relating to (a) harm/care and (b) fairness/reciprocity. Conservatives rely on these aspects but not as much. In addition they also regard as

---

\(^1\) Instituto de Física, University of São Paulo, Caixa Postal 66318, 05315-970, São Paulo - SP, Brazil
\(^2\) Complex Systems, EACH, University of São Paulo, 03828-020, São Paulo - SP, Brazil
important (c) in-group loyalty, (d) authority/respect and (e) purity/sanctity, to a larger extent than liberals.

At the scale of individuals, empirical evidence supports that: 1. moral opinions are to a large extent emitted without recourse to rules, that is, people, as a first approximation, are intuitionists; 2. there is a psychological cost of dissent, with humans trying to attain social conformity modulated by peer pressure; 3. conformity is learned from interactions within a social network; 4. individual cognitive strategies may differ with respect to the relative sensitivity to learning from novel information as compared to reinforcing habitual responses.

The notion that the methods of statistical mechanics (SM) may be employed in deriving aggregate behavior of socio-economic systems from a description of individual interactions and responses has been developed in recent years. We, however, stress that the specification of interactions between social or economic agents must be based on empirical neurocognitive and psychological evidence.

In our model of an interacting society, agents represent individuals who exchange opinions with their social neighbors about a set of issues under discussion. This information exchange defines an interaction since it causes agents to learn about other agents' opinions by changing their internal state. From MFT we extract the fact that on discussions about moral issues there are at least five dimensions to be taken into account. Thus each issue is represented by a five component unit vector $x_\mu$, (with $\mu = 1 \ldots P$). The internal state of each agent, unavailable to other agents, is given by another unit vector $J_i$ ($i = 1, \ldots, N$), also five dimensional, each component representing the weight the $i$th agent gives to each moral foundation. Unit vectors are used to avoid perceiving one agent as more moral than another.

The opinion of the $i$th agent about the $\mu$th issue is given by summing the relevance of the issue's components weighted by the strengths of the corresponding moral foundation: $h_{i\mu} = \sum_a J_{ia} x_{\mu a}$. Its attributes are its sign and magnitude, indicating respectively whether the issue is considered morally acceptable ($h_{i\mu} > 0$) or not ($h_{i\mu} < 0$) and how strongly the agent holds this position ($|h_{i\mu}|$). An opinion is formed without relying on a rule of an if-then type.

The psychological cost of disagreement between socially interacting agents $i$ and $j$ over the $\mu$th issue is quantified by $V_\delta(h_{i\mu}, h_{j\mu})$, a function of their opinions which depends on a parameter $\delta$ that measures the different treatment of corroborating or novel opinions. Conformity experiments have found that if the subject sees himself as part of a different group, the conformity effect is greatly diminished. This suggests, as a first approximation, considering the case in which agents are circumscribed to neighborhoods with homogeneous cognitive styles. The parameter $\delta$ can then be regarded as an agent specific property that we call the novelty/corroboration parameter.
Moral vectors evolve by reinforcement learning, decreasing the psychological cost under communication through a noisy channel. The social cost $H$ is defined by summing $V_\delta$ over all pairs $(i,j)$ of interacting agents and over all issues, $H = \sum_{i \neq j} V_\delta(h_{ij}, h_{ji})$. It depends on $\{J_i\}$ the configuration of the society, and on the cultural environment, given by the set of issues. Now SM permits obtaining collective or aggregate emergent phenomena arising from reinforcement learning. The value of the social cost $H$ or at least, its average, constitute relevant information to characterize the state of the society with respect to the issues under discussion. It follows that the probability of configuration $\{J_i\}$ is the Boltzmann distribution $P(\{J_i\}) \propto \exp(-\alpha H)$. The average of $H$ is fixed by the new parameter $\alpha$, which we call the peer pressure, since it sets the scale of the effect of social cost, and measures the inverse level of noise in the communication channel.

Figure 1: 1-issue psychological cost (in color in online version): $V_\delta(h_{ij}(\theta_{ij}), h_{ki}(\theta_{ki}))$ as a function of $\theta_{ij}$ and $\theta_{ki}$, the angle between $J_i$ and $J_k$ and issue $x_\mu$ where $h_{ij} = \cos \theta_{ij}$. The potential can be written as $V_\delta(h_{ij}, h_{ki}) = -h_{ij}h_{ki}$ if $h_{ij}h_{ki} < 0$ (disagreement) and $V_\delta(h_{ij}, h_{ki}) = -\delta h_{ij}h_{ki}$ if $h_{ij}h_{ki} > 0$ (agreement). For (left) $\delta = 0$, (center) $\delta = 0.4$ and (right) $\delta = 1$. The noisy learning dynamics tends to change the $J$’s making $V_\delta$ decrease along its gradient. Four peaks represent the cost of maximum disagreement when moral state vectors are opposite and angles are $(0, \pm \pi)$ and $(\pm \pi, 0)$. Note that when agents agree about the sign of their opinions, the cost of disagreement increases with $\delta$.

The functional form of the psychological cost must reflect experimental data. A reasonable choice, by no means unique, is $V_\delta(h_{ij}, h_{ji}) = \frac{1}{2}(1 - \delta)|h_{ij}h_{ji}| - \frac{1}{2}(1 + \delta)h_{ij}h_{ji}$. The corroboration/novelty parameter $\delta$ ($0 \leq \delta \leq 1$) quantifies a cognitive strategy with respect to the difference in treatment of agreement and disagreement. Our agents are conformists, namely, in the face of disagreeing opinions, the dynamics is such that the social cost is decreased. For $\delta = 0$ (figure 1a) agents are novelty seekers and do not use corroborating opinions since $V_\delta=0(h_{ij}, h_{ji})$ is flat for opinions of the same sign. For $\delta = 1$ (fig 1c) agents seek corroboration and conformity, learning equally in the case of agreement or disagreement.
We now discuss the statistical signatures that can be used to characterize the effective number of moral foundations of an agent. We compare them with equivalent signatures derived from a MFT survey\textsuperscript{23} data consisting of estimates of the moral foundation vectors extracted from questionnaires answered by a large sample of U.S. citizens ($N \approx 1.5 \times 10^4$). Subjects also declared their political affiliation ($p.a.$) with respect to social issues ranging from very liberal ($p.a.=1$) to very conservative ($p.a.=7$)\textsuperscript{14,15}.

Our main concern is the difference in the distribution of weights attributed to moral foundations by self-declared liberals and conservatives. Analytical and numerical methods\textsuperscript{27} show that the model can have two qualitatively different regimes depending on the parameters $\delta$ and $\alpha$ (figure 2). For low $\delta$ and low $\alpha$, the system is in a disordered state characterized by random correlations between the moral state vectors of the agents. Increasing either $\delta$ or $\alpha$, a transition line can be crossed into a partially ordered society. Now the agents are correlated to an average (normalized) issue $Z \propto \sum X_\mu$, the Zeitgeist vector which describes the cultural environment. The average agent is parallel to $Z$. Now we reorient $Z$, by rotating the frame of reference, so that its components are equal (e.g. $1/\sqrt{5}$ each), explicitly assuming the equivalence of all moral dimensions. Note that opinions are rotation invariant and rotating makes no numerical difference. But it does foster interpretation, since a measure of the effective number of moral foundations of agent $i$ can be defined as proportional to the sum over the moral dimensions $a_i$ of the agents moral weights: $m_{Zi} = \sum_a a_i Z_a$, the overlap between the moral vector and $Z$. 

![Figure 2](Image)

**Figure 2** The transition line separates phases with zero (below the line) and non-zero average overlap with the $Z$ vector ($\langle m_{Zi} \rangle > 0$). The phase transition is continuous. Symbols represent average and dispersion for 20 simulation runs of a $N = 400$ system with scale-free topologies and $P = 1$. The full line represents a fit to the transition border line $\alpha = k/\delta$, with $k$ constant. This can be seen more clearly in the inset. Dashed lines represent 95\% regression confidence intervals. Additional analytical justification in a mean field approximation for the border line is provided in the Supplementary Online Material.
\(Z\), ranging from -1 to 1. An agent with all moral dimensions equally important has \(m_{Zi} = 1\). Smaller values mean it relies on a reduced subset of moral dimensions. From the survey data, we extract, for each person a similar measure \(m_{Zi}\) of their number of moral dimensions. The main question is: how do the statistics of \(m_{Zi}\), from the data and from the model, compare?

Figure 3a compares histograms of \(m_{Zi}\) obtained from the data and from the model, for \(\alpha = 8\) on a scale-free social network\(^2\). But agents have no political affiliation, nor persons declare their cognitive strategy \(\delta\). The histograms permit identifying a political affiliation with a cognitive strategy, if they share a similar number of moral foundations \(\langle m_{Zi} \rangle\), independently of \(\alpha\) for broad range of \(\alpha\) (figure 3b). We establish the link: political affiliations are partially derived from subsets of moral foundations, which arise collectively from distinct cognitive strategies. We conclude that the link described in the literature connecting political affiliation to cognitive style\(^3\) arises as a consequence of social interactions.

As the order-disorder border line (figure 2) is approached from the ordered phase, the overlap with \(Z\) decreases, vanishing at the phase boundary. The best resemblance of the data and simulations occurs by identifying conservatives with agents far into the ordered phase and liberals with agents near the transition line but still in the ordered phase. Order and disorder refer to long range correlations and should not be attached to judgments of value.

We can go beyond the average number of moral foundations and look at the width of the histograms which decrease with \(\delta\) or conservative...
tendency. The same identification arises: novelty seeking behavior to liberals, corroboration to conservatives, again as a consequence of collective behavior.

Order-disorder transitions can be driven by changing the peer pressure. Even without crossing the phase boundary, the model can be used to understand collective swings from left to right, as external conditions impose increased levels of peer pressure arising from the perception of threats. The reverse swing can also be understood when conditions demand higher adaptability to new challenges. We claim that with respect to moral issues, despite the differences in opinion derived from differential reliance on moral foundations, both conservatives and liberals are on the same side of the border. Other scenarios are discernible from the phase diagram. In an application outside the realm of morality, by looking at opinions on issues for which peer pressure might be lower, a group of large $\delta$ agents, relying on corroboration, could be found in a disordered phase and seem on this set of issues, to be liberal.

This theory is semantically neutral. Evolutionary considerations should be used to dress the theory with semantics and to understand why certain foundations of morality have emerged before others and why they are different, thus breaking the remaining symmetry between the five dimensions.

We consider the most important contribution of this work to emphasize a particular methodological approach to the Social Sciences. From the description of how individuals react to incoming information obtained from social psychology empirical methods and neurocognitive data, we built an interacting model. SM leads to aggregated predictions which are tested against extensive data sets with partial information about populations. The exchange of information and the learning it elicits, induce collective emergent properties in the society not to be found in the individual. Presumably it may be useful to understand how cultural divides, such as those between conservatives and liberals, arise partly as consequences of diversity of neurocognitive mechanisms.

**METHODS SUMMARY**

In all simulations, the social neighborhood were scale-free random graphs generated by a Barabási-Albert model with branching rate $m=8$.

The transition line in figure 2 was obtained by Wang-Landau sampling of a system with hamiltonian $H$ at temperature $1/\alpha$, from the maximum of the specific heat for fixed $\delta$.

Histograms of figure 3a were obtained by Metropolis sampling. All simulations started in an ordered microstate with $\sum_{j} J_{j} J_{ka} \approx 1$.

Experimental data ($N=14250$) consisted of moral vectors with components related to five Moral Foundations in the interval [0, 5] extracted
from a specially designed questionnaire\textsuperscript{14,15}. Each vector was labeled by the subject’s self-declared political affiliation (p.a. = 1-7). We first calculated normalized moral vectors \( \{ J_i \} \) and, by defining the vector \( Z \) as the average vector within the conservative (p.a. = 6) and very conservative (p.a. = 7) classes taken together, we have calculated histograms for the effective number of moral dimensions as \( m_{Zi} = \sum J_{ia} Z_a \).

Figure 3b was prepared by calculating \( \langle m_{Zi} \rangle \), proportional to the mean effective number of moral dimensions for each p.a. class of the data and then finding, for fixed \( \alpha \), the parameter \( \delta \) that matches \( \langle m_{Zi} \rangle \) (20 Metropolis runs, \( \alpha = 6-12 \)). Only \( \alpha = 8 \) is shown Figure 3a.


27. Supplementary online material

Acknowledgements Experimental data furnished by Jonathan Haidt, whom we thank for kindly permitting us to have access to the data and also for some comments about his work. This work received support from FAPESP and CNPq under grant 550981/07-1.

Author Information Both authors contributed equally to this work. Correspondence and requests of materials should be addressed to N.C. (nestor@if.usp.br).