Universidade de São Paulo
Instituto de Física

Propriedades Coletivas Emergentes em Sociedades de Redes Neurais

Lucas Silva Simões

Orientador: Prof. Dr. Nestor Felipe Caticha Alfonso

Dissertação de mestrado apresentada ao Instituto de Física da Universidade de São Paulo, como requisito parcial para a obtenção do título de Mestre em Ciências.

Banca Examinadora:

Prof. Dr. Nestor Felipe Caticha Alfonso (Orientador - IFUSP)

Prof. Dr. Carlos Alberto de Bragança Pereira (IMEUSP)

Prof. Dr. Marco Aurelio Pires Idiart (UFRGS)

São Paulo
2018

FICHA CATALOGRÁFICA
Preparada pelo Serviço de Biblioteca e Informação
do Instituto de Física da Universidade de São Paulo

Simões, Lucas Silva

Propriedades coletivas emergentes em sociedades de redes neurais.
São Paulo, 2018.

Dissertação (Mestrado) – Universidade de São Paulo. Instituto de Física. Depto. de Física Geral.

Orientador: Prof. Dr. Nestor Felipe Caticha Alfonso.

Área de Concentração: Mecânica Estatística de Sistemas Complexos.

Unitermos: 1. Mecânica estatística; 2. Modelos de aprendizagem;
3. Entropia; 4. Teoria da informação; 5. Sociologia.

USP/IF/SBI-078/2018

University of São Paulo
Institute of Physics

Emergent Collective Properties in Societies of Neural Networks

Lucas Silva Simões

Advisor: Prof. Dr. Nestor Felipe Caticha Alfonso

Dissertation submitted to the Physics Institute of the University of São Paulo in partial fulfillment of the requirements for the degree of Master of Science.

Examining Committee:

Prof. Dr. Nestor Felipe Caticha Alfonso (Advisor - IFUSP)

Prof. Dr. Carlos Alberto de Bragança Pereira (IMEUSP)

Prof. Dr. Marco Aurelio Pires Idiart (UFRGS)

São Paulo
2018

Acknowledgements

First, I acknowledge the financial support of CNPq for fellowship grant *n*º134812/2016 – 6 and FAPESP for fellowship grant *n*º2016/15860 – 3. In times when scientific research has lacked the proper sponsorship from the Brazilian governors, to be awarded with that support and trust is a real blessing.

Thanks Nestor for the inspiration throughout what has been my scientific career so far. Thanks Felipe, André e André for all the discussions of the groups' work and for the company. It has been much fun.

Thanks Renzo, Paulo, Luiz Felipe, Gabriel, Rodrigo, Artur, Estevão, Felipe, Rodolfo, Thiago, Ramon, Nickolas, Patrick and many other people from the Ciências Moleculares for the great years we spent together at USP.

Special thanks goes to my family: my mother Adriana, my father Márcio and my brother Gabriel. Always backing up my choices and providing a comforting environment at home, all while mocking me and my “unusual” line of research. I would not expect less from you. And also my best friend and girlfriend, Giuliana, without whom these past years maybe would be bearable, but certainly would be without color and much much more difficult. To you four my deepest thanks.

“Vanity of vanities and all is vanity (Ecclesiastes 1, 2), except to love God and serve Him alone.”¹. Ultimately this is all to Him. Thanks would be too little.

¹The Imitation of Christ, chapter I

Abstract

This project deals with the study of the social learning dynamics of agents in a society. For that we employ techniques from statistical mechanics, machine learning and probability theory.

Agents interact in pairs by exchanging for/against opinions about issues using an algorithm constrained by available information. Making use of a maximum entropy analysis one can describe the interacting pair as a dynamics along the gradient of the logarithm of the evidence. This permits introducing energy like quantities and approximate global Hamiltonians. We test different hypothesis having in mind the limitations and advantages of each one. Knowledge of the expected value of the Hamiltonian is relevant information for the state of the society, inducing a canonical distribution by maximum entropy. The results are interpreted with the usual tools from statistical mechanics and thermodynamics.

Some of the questions we discuss are: the existence of phase transitions separating ordered and disordered phases depending on the society parameters; how the issue being discussed by the agents influences the outcomes of the discussion, and how this reflects on the overall organization of the group; and the possible different interactions between opposing parties, and to which extent disagreement affects the cohesiveness of the society.

keywords: Statistical Mechanics, Learning Models, Entropy, Information Theory, Sociology

Resumo

Esse projeto lida com o estudo da dinâmica de aprendizado social de agentes em uma sociedade. Para isso empregamos técnicas de mecânica estatística, aprendizado de máquina e teoria de probabilidades.

Agentes interagem em pares trocando opiniões pró/contra questões usando um algoritmo restringido pela informação disponível. Fazendo-se uso de uma análise de máxima entropia, pode-se descrever o par da interação como uma dinâmica ao longo do gradiente do logaritmo da evidência. Isso permite introduzir quantidades similares a energia e Hamiltonianos globais aproximados. Testamos diferentes hipóteses tendo em mente as limitações e as vantagens de cada uma. Conhecimento do valor esperado do Hamiltoniano é informação relevante para o estado da sociedade, induzindo uma distribuição canônica a partir de máxima entropia. Os resultados são interpretados com as ferramentas usuais de mecânica estatística e termodinâmica.

Algumas das questões que discutimos são: a existência de transições de fase separando fases ordenada e desordenada dependendo dos parâmetros da sociedade; o como a questão sendo discutida pelos agentes influencia os resultados da discussão, e como isso se reflete na organização do grupo como um todo; e as possíveis diferentes interações entre partidos opostos, e até que ponto o desacordo afeta a coesão da sociedade.

palavras-chave: Mecânica Estatística, Modelos de Aprendizagem, Entropia, Teoria da Informação, Sociologia

Table of Contents

1	Introduction	6
1.1	Outline	8
2	Background	9
2.1	Social Influence and Change	9
2.1.1	Psychology: Conformity and Group Categorization	10
2.1.2	Neuroscience: Social Exclusion and Error Correcting	11
2.2	Moral Foundations Theory	12
2.3	Probability Theory	13
2.4	Entropy and Inference	15
2.4.1	Entropy as an Inference Tool	15
2.4.2	Statistical Mechanics from Maximum Entropy	16
2.5	Entropic Dynamics	18
2.5.1	The basics	19
2.5.2	Agent model definition	22
2.5.3	Calculating Z_{n+1}	26
3	Models	28
3.1	Model 1	29
3.2	Model 2	33
3.3	Model 3	35
3.3.1	Infinite-range Ising model	36
3.3.2	Bipartite Society	38
4	Results	42
4.1	Model 1	42
4.2	Model 2	46
4.3	Model 3	49

<i>TABLE OF CONTENTS</i>	3
5 Conclusion and Perspectives	54
5.1 Discussion of the results	54
5.2 Final Remarks	57
References	59

List of Figures

2.1	Schematic representation of the update procedure done to revise the distribution Q_n . It goes as follows: instead of using Bayes' Theorem (blue path), one updates the distribution through Maximum Entropy (red path) while retaining the expected values matching with the ones of Bayes.	20
3.1	Exemplification of the interactions between and inside the 2 communities in a bipartite society	39
4.1	Solutions of Equation 3.16. Top: The normalization of the MF distribution ζ as a function of the social pressure and number of neighbors ($\beta\nu$). Bottom: Magnetization m . The other order parameter r has a similar behavior to m	43
4.2	Phase diagram in the space $\gamma \times \beta\nu$. A phase transition separates an ordered from a disordered phase as signaled by the value of the order parameter m . Here, the value of ε was 0.1 and as it grows towards 0.5 the ordered phase decreases.	44
4.3	Modulation function/gradient of the evidence for different values of γ . The noise ε is fixed at 0.2, but changing to other values between 0. and 0.5 does not change qualitatively the figure	44
4.4	Probability distribution $P(h \varepsilon, \gamma, \beta\nu)$ for a set of values for the model's parameters.	45
4.5	Variance of the field h with respect to selected values of $\beta\nu$ and γ . . .	45
4.6	The 4 different Zeitgeist hypothesis shown for comparison. All extracted from MFQ's respondents data.	47

4.7	Different posteriors for $\tilde{\beta}$ given the chosen Zeitgeist and the political affiliation of the responses $\{h\}$. Since the distributions are sharp, the error bars in the y -axis fall inside the markers (which are centered around the mean values)	47
4.8	The histogram of opinions h for a given pa group considering an specific Zeitgeist and the corresponding best fit of the model $P(h \tilde{\beta})$ given the data	48
4.9	The phase diagram when $\beta = 10.0$ and $N_A = N_B$. The initial points for the iterative algorithm were $x_0 = 0.9, y_0 = 0.1$	51
4.10	Free energy $f(x, y)$ landscape when $\beta = 10.0, \delta = 0.25, \varepsilon = 0.6$ and $N_A = N_B$. This is a region in which there is no consensus among the communities	52
4.11	Free energy $f(x, y)$ landscape when $\beta = 10.0, \delta = 0.8, \varepsilon = 0.2$ and $N_A = N_B$. This is a region in which there is overall consensus among the society	52
4.12	Free energy $f(x, y)$ landscape when $\beta = 10.0, \delta = 0.8, \varepsilon = 0.8$ and $N_A = N_B$. This is a region in which there is internal consensus on each community but they disagree with one another.	53

1

Introduction

How can we integrate the fact that sometimes people agree on their moral valuations with one another, and other times disagree? One can naturally imagine two alternative situations: the first one, in which the whole society agrees on what is considered correct, and another one in which chaos reign, because it cannot seem to reach a consensus. How is it then that our societies maintain a considerable level of disagreement while still retaining some organization? How can there be a coexistence of a variability of opinions and still there be a cohesiveness that enables a society to thrive and develop? In times when we are haunted with cultural and political wars between people with different opinions, one cannot but wonder which are the relevant variables that describe this cohesiveness and richness of cultures in a society, and what one can expect about times to come.

Several people have investigated those questions, each with their different background and perspective into the difficult problem of understanding human organization in society. Those people came most notably from areas such as sociology, anthropology and psychology, but more recently also physics and neuroscience. We present some of the ones that influenced the research shown in this dissertation.

On the social sciences, one of those works we build solidly upon is the line of research from Jonathan Haidt and collaborators on social psychology of morality (see e.g. Haidt 2001, 2007; Haidt and Kesebir 2010).

It is also important to consider existing experiments and models from neurophysiology and cognitive neuroscience, specially findings about what drives human behaviour or how to categorize and understand human learning.

Both areas - social psychology and neuroscience - are deeply related, to the extent that (Greene 2009) says: “[the field of cognitive neuroscience of morality] provides a set of useful entry points into the broader problems of complex cognition and decision making”. If one wants to understand the brain and human behavior, one ought to attack this problem knowing both endeavours.

For the past few decades we have seen several quantitative contributions to approach the social sciences. A constant throughout the majority of the enterprises has been the use of tools and ideas from physics (most specially statistical mechanics). This has led some to name this variety of concepts and results as *sociophysics*. One can check (Castellano, Fortunato, and Loreto 2009) for a review of those.

Here we mention some of the ideas that influenced the field and our research:

- Schelling’s models for segregation (Schelling 1969, 1971);
- Axelrod’s models for dissemination of culture (Axelrod 1997) and for the evolution of cooperation (Axelrod and Hamilton 1981);
- Caticha and Vicente’s models on social consensus (Caticha and Vicente 2011), morality and political affiliation descriptions (Vicente et al. 2014; Caticha, Cesar, and Vicente 2015) and community formation (Alves and Caticha 2016)

This dissertation fits in the line of work from this last point. The tools used are those of bayesian and entropic inference, statistical physics and machine learning. Bayesian and entropic learning because we are dealing with situations of incomplete information; statistical mechanics because it is a natural extension of the first two and provides a toolbox of methods developed by physicists throughout the years (such as mean field, Laplace’s method to find a partition function, among others) to solve the problems encountered; and machine learning to inform our choice of models and our understanding of learning from a quantitative perspective.

As mentioned at the beginning of this section, we are interested in the interplay between aggregation and dispersion forces in society, and how this gives rise to the social and cultural diversity we find in our societies. This being a broad question, we narrow it down to some topics:

- first, we describe a model for agents learning from their peers using entropic inference, and propose a mean field analysis for a society of this agents. This model will give us interesting results, such as a phase transition between ordered and disordered phase in the organization of the opinions of the agents, but it has many extraneous features that hinder our analysis with available data from Haidt and collaborators;
- then, we proceed to a simpler model, which still captures the relevant characteristics of the first model, while being more manageable to compare with experimental data.

This will give us more intuition into the kind of predictions we can make about agents and their organization in communities (more specifically, in the data, in political affiliations). The most relevant characterizations we make are of extremists agents and of the choice of the discussed question in the society;

- finally, we proceed to a model which assumes the existence of two adversarial communities. Here we are interested in the coexistence of those communities in face of conflict and delimitation of different “cultures”.

Since this line of research is pioneering, the number of questions abound and many times it is difficult to narrow down specific motives and hypothesis. A good research, in that scenario, may not always answer a comprehensible and well delimited question about the system being modeled, but instead leverage understanding in what variables and descriptions better capture the phenomena under observation. For that reason, although we do not presume to be called giants, we believe that sometimes one must do this “foraging”¹ task to support future work of greater minds.²

1.1 Outline

Section 2 provides an introductory background to the topics relevant to this study. We start talking about the studies in psychology and neuroscience that form the substrate of our investigation and give support to our hypothesis. Then we present the quantitative tools we use in our study, namely: probability theory, entropic inference and statistical physics. Finally, we explain the reasons we have to believe that questions usually called “social sciences’ problem” can be posed, using the tools presented, as a physics problem.

In Section 3 we present three different models that we constructed to investigate our hypothesis, presenting the results in Section 4 and the discussion in Section 5.

¹“I can see no other escape from this dilemma (lest our true aim be lost for ever) than that some of us should venture to embark on a synthesis of facts and theories, albeit with second-hand and incomplete knowledge of some of them, and at the risk of making fool of ourselves. So much for my apology.” (Schrödinger 1944)

²“If I have seen further it is by standing on the shoulders of Giants.” (Newton 1675)

2

Background

Here we describe the main tools and ideas being used to study the questions proposed in Section 1.

As told, there has been many enterprises to study the mentioned problems and also others related. Some of the research areas that are more aligned with our interests are those of (social) psychology and (cognitive) neuroscience. We also present some developments in bayesian inference (probability theory) and entropic inference, which form the solid ground of our models, together with statistical mechanics and some concepts from machine learning.

2.1 Social Influence and Change

Human beings are social (Baumeister and Leary 1995) and adaptive beings relying on several automatic and intuitive neural mechanisms to change their behaviours.

Here we show some psychological and neurological evidence to support that claim. This is important because it give us reason to believe that we can model, at least to some extent, people's interaction in society and their learning from peers and mistakes.

2.1.1 Psychology: Conformity and Group Categorization

For many the pioneer of social psychology experiments, Sherif dedicated himself to the study of *attitudes* and *social influence*. In particular, one of his first articles studied what he called “the daily phenomena of suggestion (...) in the formation of attitudes” (Sherif 1937, pg. 98). On that experiment participants had to perform the same visual task twice: at the first trial they did it in pairs; and in a second trial, individually.

What Sherif noted was twofold:

- During the first trial, there is a norm (standard) and a range of responses that are established within the group, and subjects typically update their responses to conform to the group;
- This conformity is retained when the subject comes back to the second trial, several days later, showing that this kind of influence has lasting effects on the subject.

So far several clarifications and further results have been obtained as well. We present some of those for completeness of this exposition.

First, (Deutsch and Gerard 1955) makes a distinction between 2 kinds of influences: “*normative*” (“influence to conform with the positive expectations of another”) and “*informational*” (“an influence to accept information obtained from another as evidence about reality”). This distinction is important because we can have settings in which subjects do not feel as belonging to a particular group - therefore they are not subject to a normative influence - and yet they can be influenced in an “informational” way when they consider their peers to be a trustworthy source of information about reality.

Around the same time, (Asch 1956) set out to study the effect of *majorities* on the opinion of other participants of a group. They performed a visual task with groups of 7-9 individuals, of which only one had not met previously with the experimenter. Those, called confederates, were set respond the task with a previously established strategy: the idea was that the majority could, after setting a baseline of trust with correct answers, dissuade the subject to give an incorrect answer. The results obtained were mixed: as Asch puts it, “the performances ranged from complete independence to complete acquiescence”, and “one-third of the minority estimates were distorted toward the majority”.

To give one last perspective, (Abrams et al. 1990) gathered experimental setups from several authors (Sherif and Asch among them) and studied those setups giving a closer look to *categorization* (of self and of others). Their study showed how the membership of some subject’s peers modulated the influence they had on the subject’s response. Amidst other things, they concluded that, for both informational and normative influences, *self-categorization* had a crucial contribution to the conformity behaviour. In other words: defining the group to which I belong and realizing my peers belong to the same group (or do not) are necessary

to make me trust them and take their opinions for granted (or not).

2.1.2 Neuroscience: Social Exclusion and Error Correcting

We also have evidence from neuroscience (Eisenberger, Lieberman, and Williams 2003) explaining how is it that *social pain*¹ experiences recruit the same neural machinery employed during physical pain ones.

During their experiment, several participants had to play a virtual game in which they experienced being socially rejected by other peers in 2 different situations: first in an implicit event and then an explicit exclusion scenario. In both cases the researchers found a distinguished fMRI (functional magnetic resonance imaging) signal in two regions of the subjects' brains: dorsal anterior cingulate cortex (*dACC*) and right ventral prefrontal cortex (*RVPFC*).

In fact, growing number of research is surfacing the importance of this interplay between physical and social distress. (Eisenberger and Lieberman 2004) says that social support in humans even "increases tolerance to electric shock stimulation (Buck and Parke 1972) and decreases levels of self-repeated pain during cold-pressor task (Brown et al. 2003)". A recent review by (Eisenberger 2012) analyses those claims in more depth and has pointers to further evidence for the interested reader.

On another account, researchers have observed (Gehring et al. 1993) there was a neural signal consistently present after subjects committed errors in a choice task. This signal was dubbed error-related negativity (*ERN*). There have been extensive experimental (Braver et al. 2001) and theoretical (Holroyd and Coles 2002; Yeung, Botvinick, and Cohen 2004) study to understand where the signal comes from and to model the behaviour of the system in a comprehensible way. Some of the results obtained by the references above show that the *ACC* is one of the brain areas essential to generate these *ERN* signals.

For historical accounts of the development of these ideas (and also great reviews), one can see (Gehring et al. 2018; Dehaene 2018).

We present these studies on brain's error-correcting mechanisms in this subsection because there is a connect between them and the studies of social exclusion: both have the *ACC* in common and both deal with situations in which the human brain is dedicated to changing one's perspective and adapt in face of error or discomfort.

The evidence presented at this section motivates our understanding that humans can be described - to some level and at some tasks - as agents updating their beliefs and behaviours

¹"Social pain is defined as the distressing experience arising from the perception of actual or potential psychological distance from close others or a social group" (Eisenberger and Lieberman 2004)

to conform to their peers. Either because they feel pressured to conform with the group's moral norms and cultural views, or because they feel distress from being rejected out of the group (because they did not conform before). Some of these results will also illuminate our conclusions after we present the results from the models we propose.

2.2 Moral Foundations Theory

How do people assess morality in everyday life? Recently this has been investigated extensively by Haidt and collaborators (a nice explanation can be found in Haidt 2007). We outline some of their conclusions through several years and articles below.

In (Haidt 2001) was proposed that moral judgment is not a consequence of moral reasoning, but of “quick, automatic evaluations (*intuitions*)”. According to this proposal moral reasoning “is an effortful process, engaged in after a moral judgment is made, in which a person searches for arguments that will support an already-made judgment” (pg. 818)

Although there are well-stated criticisms (Pizarro and Bloom 2003) toward Haidt's *moral intuitionist* model², it is already clear that emotions and intuitions are primordial when humans make moral judgments (Greene 2009). This is important to us because it is easier to model intuitive mechanisms of the brain than more elaborate ones. This also inform us that the human's action and reaction to moral assertions is prompt, immediate and mostly unconscious.

Also from this line of research, they found support for the idea that the human mind comes equipped with 5 “modules” (Haidt and Joseph 2004)³ to judge morality: Harm, Fairness (the ones agreed upon throughout academic community), Loyalty, Ingroup/ Authority and Purity/Sanctity. This is now known as the *Moral Foundations Theory* (MFT Graham et al. 2013).

This is a core pillar to our research because it is what guarantees that the assessment of morality happens within a finite set of dimensions, even if psychologists happen to find more foundations later. Morality does not happen in a non-comprehensible infinite space, nor is it an intermingle of “yes or no” output to one-case judgments.

Finally, (Gilligan et al. 2012) developed a “reliable and theoretically-grounded measurement of the full range of moral concerns”, the *Moral Foundations Questionnaire* (MFQ).

²Namely, they argue for the relevance of some moral reasoning mechanism apart from the intuition circuitry. In their words, “there is considerable evidence from outside the laboratory that people actively engage in reasoning when faced with real-world moral dilemmas. Together, these facts limit the strong claims of the social intuitionist model concerning the irrelevance of conscious deliberation.”

³At the time they thought to be only 4. Now there is a discussion to propose a sixth one: maybe a Liberty dimension should be included to account for the libertarian movement (Iyer et al. 2012)

It is set of responses people gave stating their opinions on 30 questions about morality.⁴ Those questions ranged through all of the moral foundations developed by the authors in previous works. Subjects also labeled themselves with what the researchers called **political affiliations** (*pa*), which could one of the following entries: Very Liberal, Liberal, Slightly Liberal, Moderate, Slightly Conservative, Conservative, Very Conservative, Libertarian, Other and Don't know (we disregarded the last 3 categories in our analysis in Section 4.2).

This piece of their research culminated in discovering that people from different political affiliations tend to see moral assertions through different “lenses”: people more inclined to be defined as Liberals tend to give more importance to features such as Harm and Care, whereas people self-labeled as Conservatives aggregate Ingroup, Loyalty and Purity dimensions to the first two (we reproduced this result from the MFQ's data in Figure 4.6).

This is an important piece of information for social psychology, a field which historically has focused much on Harm and Care, but has not had many accompanying developments on the study of the other foundations.

2.3 Probability Theory

Probability theory is often an underestimated tool in a scientist career. Much of the failure to see the importance of probability theory comes from an incorrect understanding of what it is.⁵ Therefore we present the outline of a derivation of the rules of probability and explain what they are meant to be and what one is meant to do with them.⁶

We begin by defining the problem at hand: we want a mathematical tool to reason in situations of incomplete information and that is an extension to Boolean logic. That is, we want a tool that incorporate the concepts of **True**, **False**, **and**, **or**, **not** and extend them to deal with inference settings in which we cannot be certain about the veracity or falsehood of assertions $a, b, \dots$; we must be able to say whether some assertion is more “plausible” $\succ$ than another one or not.

Another necessary characteristic is that this tool has to work by default with variable contexts of information, otherwise we could not talk about “incomplete information”. We denote this idea of “background information” by the backlash symbol “|”, e.g. “ $a|c$ ” means “consider the plausibility of assertion a given that we know c is true”.

⁴One can take this same test and evaluate oneself at <https://www.yourmorals.org/>

⁵We agree with the vision Laplace had: “Probability theory is nothing but common sense reduced to calculation” (Sivia and Skilling 1998)

⁶We follow mostly the expositions by (Caticha 2012), built on work by (Cox 2001; Jaynes and Bretthorst 2003 and several others)

The requirements above are only the ground construction of our “plausibilities” theory. So far we could have many tools fulfilling those. That is why we constrain our theory with a set of *desiderata* (desires) that our method must also obey:

1. We want “plausibilities” to be *transitive*, that is, if $a|b \succ a|c$ and $a|c \succ a|d$, then it must follow that $a|b \succ a|d$;
2. We want that for every possible way of composing a given assertion the plausibilities of them all must be the same;
3. There must be 2 numbers v_T and v_F , at first unknown, that represent the concepts of **True** and **False**, respectively; that is, $a|a = v_T$ and $a|\text{not } a = v_F$;
4. There must be 2 functions F and G , at first unknown, to represent the logical operators **and** and **or**.

As much as it may sound absurd, it can be proved (as is in fact done in e.g. Cox 2001; Jaynes and Bretthorst 2003; Caticha 2012), that the rules of probability theory follow from these 4 *desiderata*.

To be more explicit, the rules are as follows:

$$p(a|a) = v_T = 1, \quad p(a|\bar{a}) = v_F = 0 \quad (2.1)$$

$$p(a + b|c) = p(a|c) + p(b|c) - p(ab|c) \quad (2.2)$$

$$p(ab|c) = p(a|c)p(b|ac) = p(b|c)p(a|bc) = p(ab|c) \quad (2.3)$$

$$p(a|c) + p(\bar{a}|c) = 1 \quad (2.4)$$

and from these simple rules we can obtain all the other known relations for probabilities.

The advantage of the method we outlined is that, instead of describing a mathematical tool first and then trying to interpret and understand its function, we start by defining the purpose of our tool and from that the functioning comes naturally.

Now that we know that probabilities are the mathematical object describing plausibilities of assertions in face of incomplete knowledge, one can develop other insights using them.

For example, one can take the product rule (Equation 2.3) to prove the Bayes’ rule:

$$p(b|ac) = \frac{p(a|bc)p(b|c)}{p(a|c)} = \frac{p(a|bc)p(b|c)}{\sum_b p(a|bc)p(b|c)} \quad (2.5)$$

Let us illustrate: consider b = “it will rain today”, the assertion we want to reason about; c = “I am in São Paulo in January, a city where there is on average 15 rainy days in that time

of the year”⁷, our background information; and a = “there are dark clouds in the sky”, the new experimental data to which we have access. In that informational scenario, a person using Equation 2.5 most certainly would not leave the house without an umbrella. That is because the new data a (provided we have background information c), produces further insight into assertion b , rendering $p(b|ac) > p(b|c)$.

This understanding of probability theory will permeate our assumptions and developments throughout this dissertation, and we also employ Bayes’ rule in Sections 2.5, 3.2.

2.4 Entropy and Inference

Entropy has always been an elusive concept to physicists (Jaynes 1965, 1980). Starting with (Jaynes 1957a, 1957b) an understanding of entropy as a tool for making inferences about the world has began. After some developments (Shore and Johnson 1980; Skilling 1988) in the theory of Maximum Entropy (ME), now we seem to be reaching a point in which researchers can derive more complex physics theories as an application of the ME formalism (Caticha 2017).

In this section and in Section 2.5 we delineate the importance of that method and the results we obtain with it to the specific case of social agents.

2.4.1 Entropy as an Inference Tool

The method to find the proper tool to update our inferences under situations of incomplete information will be the same as done with probabilities in Section 2.3: we start describing the problem at hand, we impose some *desiderata* to our method, and then try to prove the existence of some method abiding by those *desiderata*. Luckily this path was already staged by many before us on the references given before, so we only describe the constraints desired and the tool that follows.

1. We want the method to be universal;
2. We want the update to be minimalist;
3. We want locality in the space of configurations;
4. We want invariance of labels (change of variables);
5. We want independent systems to be describes evenly together or separately

We do not present the full derivation due to its length and it not being the focus of this dissertation. One can look at (Caticha 2012) for a pedagogical exposition.

⁷Taken from https://en.wikipedia.org/wiki/S%C3%A3o_paulo#Climate

The method obtained is:

In order to *update* from a probability distribution $q(\mathbf{x})$ when new information arrives and forces a review of beliefs, one must choose the distribution $p(\mathbf{x})$ that maximizes the functional

$$S[p|q] = - \int d\mathbf{x} p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) \quad (2.6)$$

subject to the constraints imposed by new information.

Note there is a philosophical statement when we say that ME is a method of *updating* probability distributions, or beliefs. Doing that we exclude the possibility of having a situation of “information emptiness”. There is always a background, a context, over which we perform our inference. This is a non-trivial but central characteristic of the method developed.

2.4.2 Statistical Mechanics from Maximum Entropy

Now, with ME at our disposal, one can obtain several interesting results as corollaries to different informational settings. We show one for its relevance in physics: *Statistical Mechanics*.

We must first define what is our informational context and space of discourse, then we must impose the constraints to our update and this will lead us to the ME theory relevant to our inference.

In the case of Statistical Mechanics we consider the space over which our inference is made as the phase space $\mathbf{x} = (\mathbf{q}, \mathbf{p})$. Also, (Caticha 2012) provides good arguments based in Hamiltonian dynamics to propose that the prior distribution $q(\mathbf{x})$ must be an uniform distribution over the phase space.

Now we can proceed with the calculation of what is called the family of *canonical distributions*. Consider a set of functions $f_a(\mathbf{x})$ known to provide relevant about the system one wants to study, e. g. a Hamiltonian. We happen to know the expected values of those functions: $\langle f_a(\mathbf{x}) \rangle = F_a$, e.g. we perform a measurement exactly because we know those quantities are important to describe the problem. Maximum Entropy then tells us that we must maximize the functional S with respect to posterior distribution p taking the information on the expected values in consideration:

$$\mathcal{L}[p|q] = - \int d\mathbf{x} p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) + \alpha \left(\int d\mathbf{x} p(\mathbf{x}) - 1 \right) + \lambda^a \left(\int d\mathbf{x} p(\mathbf{x}) f_a(\mathbf{x}) - F_a \right) \quad (2.7)$$

where the constraint for α comes from the fact that probability distributions must be properly normalized.

Variational calculus arguments lead us into

$$p(\mathbf{x}) = \exp(-1 - \alpha - \lambda^a f_a(\mathbf{x})) \quad (2.8)$$

The Lagrange multipliers $\alpha, \{\lambda_k\}$ still need to be calculated taking in account the constraints imposed. We proceed with the normalization constraint:

$$1 = \int d\mathbf{x} p(\mathbf{x}) = \int d\mathbf{x} \exp(-1 - \alpha - \lambda^a f_a(\mathbf{x})) \quad (2.9)$$

$$= \exp(-1 - \alpha) \int d\mathbf{x} \exp(-\lambda^a f_a(\mathbf{x})) \quad (2.10)$$

From which we obtain the definition of a *partition function*, or the normalization, of our distribution:

$$Z(\lambda) \equiv \exp(1 + \alpha) = \int d\mathbf{x} \exp(-\lambda^a f_a(\mathbf{x})) \quad (2.11)$$

and now we can rewrite Equation 2.8 in a more familiar way:

$$p(\mathbf{x}) = \frac{1}{Z} \exp(-\lambda^a f_a(\mathbf{x})) \quad (2.12)$$

For the other constraints, we do:

$$F_a = \langle f_a \rangle = \frac{1}{Z} \int d\mathbf{x} f_a \exp(-\lambda^a f_a(\mathbf{x})) \quad (2.13)$$

$$= \frac{1}{Z} \int d\mathbf{x} \left(-\frac{\partial}{\partial \lambda^a} \right) \exp(-\lambda^a f_a(\mathbf{x})) = \frac{1}{Z} \left(-\frac{\partial}{\partial \lambda^a} \right) Z(\lambda) \quad (2.14)$$

This lead us into the following relation:

$$F_a = -\frac{\partial}{\partial \lambda^a} \log Z \quad (2.15)$$

This is the famous relation from statistical mechanics relating the derivatives of the partition function with intensive variables, such as E .

A final result we want to expose about canonical distributions is the relation we obtain when we insert the maximum entropy distribution p back into the definition of entropy S (Equation 2.6, remembering the uniform prior):

$$S_{\text{MAX}} = - \int d\mathbf{x} p(\mathbf{x}) \log p(\mathbf{x}) \quad (2.16)$$

$$= - \int d\mathbf{x} \frac{1}{Z} \exp(-\lambda^a f_a(\mathbf{x})) [-\log Z - \lambda^a f_a(\mathbf{x})] \quad (2.17)$$

Therefore, we find the usual Legendre transform used in statistical mechanics.

$$S(F) = \log Z(\lambda) + \lambda \cdot F \quad (2.18)$$

We have done the calculations for a generic distribution from the canonical family. Particularizing to the typical case in which one uses only information regarding the Hamiltonian $\langle \mathcal{H} \rangle = E$, the results above translate into:

$$p(\{q_i, p_i\}) = \frac{1}{Z} \exp(-\beta \mathcal{H}(\{q_i, p_i\})) \quad (2.19)$$

$$Z = \int \prod_i dq_i dp_i \exp(-\beta \mathcal{H}(\{q_i, p_i\})) \quad (2.20)$$

$$E = -\frac{\partial}{\partial \beta} \log Z \quad S = \log Z + \beta E \quad (2.21)$$

which one can recognize as what is usually called *Canonical Ensemble* in physics.

2.5 Entropic Dynamics

The idea of using entropy to update probability distributions consistently and sequentially, called henceforth *Entropic Dynamics*, has proven to be a powerful concept in physics and in other areas as well.

In our case, we follow (Cesar 2014; Alves 2015; Alves and Caticha 2016, 2018) and use Entropic Dynamics to describe a model for social agents. At first these agents will just be machines learning how to classify vectors into two groups in the presence of multiplicative noise, but we imbue them with a semantical meaning in Section 2.5.2.

2.5.1 The basics

Incomplete knowledge about the variables $\mathbf{x}$ under study is represented by distribution $Q(\mathbf{x})$, which belongs to a family defined by a set of generator functions $\{f^a\}_{a=1}^G$. Once the expected values $\langle f^a \rangle = \eta^a$ are known, maximum entropy yields the distribution with least biased information at a given timestep n is:

$$Q_n(\mathbf{x}) = \frac{1}{\zeta_n} \exp \left(- \sum_{a=1}^G \lambda_n^a f^a(\mathbf{x}) \right) \equiv P(\mathbf{x}|\lambda_n) \quad (2.22)$$

where $\lambda_n = \{\lambda_n^a\}$ are the Lagrange multipliers that enforce the constraints chosen at timestep n .

At timestep $n + 1$ new information y_{n+1} is discovered and we update the inference $Q_n \rightarrow Q_{n+1}$ using Maximum Entropy. This will be the Entropic Dynamics ruling the learning of our agents.

Before calculating the update, we prove some identities that are going to be useful later on⁸:

$$\begin{aligned} \frac{\partial \log \zeta}{\partial \lambda_b} &= \frac{1}{\zeta} \frac{\partial \zeta}{\partial \lambda_b} = \frac{1}{\zeta} \int d\mathbf{x} (-f^b(\mathbf{x})) e^{-\lambda_a f^a(\mathbf{x})} \\ \therefore \frac{\partial \log \zeta_n}{\partial \lambda_b^n} &= -\eta_n^b \end{aligned} \quad (2.24)$$

$$\begin{aligned} \frac{\partial Q}{\partial \lambda_b} &= \frac{1}{\zeta} e^{-\lambda_a f^a(\mathbf{x})} (-f^b(\mathbf{x})) + e^{-\lambda_a f^a(\mathbf{x})} \left(\frac{-1}{\zeta^2} \right) \frac{\partial \zeta}{\partial \lambda_b} \\ \therefore \frac{\partial Q}{\partial \lambda_b} &= [\eta^b - f^b(\mathbf{x})] Q \end{aligned} \quad (2.25)$$

In possession of the prior distribution $Q_n(\mathbf{x})$ and a model for the generation of data $P(y_{n+1}|\mathbf{x})$, the best inferential update when new data y_{n+1} becomes available is Bayes' Theorem:

$$P_{n+1}(\mathbf{x}) = P(\mathbf{x}|y_{n+1}) = \frac{L_{n+1} Q_n}{Z_{n+1}} = \frac{P(y_{n+1}|\mathbf{x}) P(\mathbf{x}|\lambda_n)}{\int d\mathbf{x} P(y_{n+1}|\mathbf{x}) P(\mathbf{x}|\lambda_n)} \quad (2.26)$$

⁸Here we start using Einstein summation convention

$$x_a y^a \equiv \sum_a x_a y_a \quad (2.23)$$

This update has the problem of the new distribution P_{n+1} not always belonging to the parametric family initially chosen. We need an alternative step that retains the relevant information from Bayes' rule while also ensuring we obtain the same functional form as the prior. This procedure is illustrated in Figure 2.1.

Figure 2.1: Schematic representation of the update procedure done to revise the distribution Q_n . It goes as follows: instead of using Bayes' Theorem (blue path), one updates the distribution through Maximum Entropy (red path) while retaining the expected values matching with the ones of Bayes.

To do that we maximize the entropy $S[Q_{n+1}||Q_n]$ while enforcing that the new expectation values match that of P_n :

$$\begin{aligned} \mathcal{L}[P, Q, \{\Delta_a\}] = & \int d\mathbf{x} Q_{n+1} \log \frac{Q_{n+1}}{Q_n} - \Delta_0 \left[\int d\mathbf{x} Q_{n+1} - 1 \right] \\ & - \Delta_a \left[\int d\mathbf{x} f^a Q_{n+1} - \mathbb{E}_{P_{n+1}}[f^a(\mathbf{x})] \right] \end{aligned} \quad (2.27)$$

Being both Q_n and P_n already fixed, we can only vary the posterior distribution Q_{n+1} . Taking a functional derivative with respect to it, we find:

$$\delta \mathcal{L} = \int d\mathbf{x} \delta Q_{n+1} [\log Q_{n+1} + 1 - \log Q_n - \Delta_0 - \Delta_a f^a] \quad (2.28)$$

Equating this to zero, we obtain the expression for the Maximum Entropy (ME) posterior:

$$Q_{n+1}(\mathbf{x}) = Q_n(\mathbf{x}) e^{-1+\Delta_0+\Delta_a f^a} = \frac{1}{\zeta_{n+1}} \exp(-\lambda_a^{n+1} f^a(\mathbf{x})) \quad (2.29)$$

where ζ_{n+1} is the new normalization factor, λ_a^{n+1} is defined as $\lambda_a^n + \Delta_a$, and we finally get the updated distribution at timestep $n + 1$.

Taking a derivative with respect to Δ_b in Equation 2.27 the constraint adopted becomes explicit:

$$\mathbb{E}_{P_{n+1}}[f^b(\mathbf{x})] = \mathbb{E}_{Q_{n+1}}[f^b(\mathbf{x})] \equiv \eta_{n+1}^b \quad (2.30)$$

Subtracting η_n^b from both sides and working out the expression with Equation 2.24 and Equation 2.25, we find an update rule for the parameters of the distribution.

$$\eta_{n+1}^b - \eta_n^b = \mathbb{E}_{P_{n+1}}[f^b(\mathbf{x})] - \eta_n^b = \int d\mathbf{x} f^b(\mathbf{x}) P_{n+1} - \eta_n^b \int d\mathbf{x} P_{n+1} \quad (2.31)$$

$$= \int d\mathbf{x} [f^b(\mathbf{x}) - \eta_n^b] P_{n+1} = \int d\mathbf{x} \frac{L_{n+1}}{Z_{n+1}} [f^b(\mathbf{x}) - \eta_n^b] Q_n \quad (2.32)$$

$$= \int d\mathbf{x} \frac{L_{n+1}}{Z_{n+1}} \left(-\frac{\partial Q_n}{\partial \lambda_n^b} \right) = -\frac{1}{Z_{n+1}} \frac{\partial}{\partial \lambda_n^b} \left(\int d\mathbf{x} L_{n+1} Q_n \right) \quad (2.33)$$

Finally, the schematics represented by Figure 2.1 can be described as one simple a gradient descent evolution:

$$\eta_{n+1}^b = \eta_n^b - \frac{\partial}{\partial \lambda_n^b} \log Z_{n+1} \quad (2.34)$$

where $Z_{n+1} = \int d\mathbf{x} P(y_{n+1}|\mathbf{x})Q_n(\mathbf{x})$ is the evidence of the model, and the distribution $P(y_{n+1}|\mathbf{x})$ the likelihood, which describes how in our inference the data y_n ought to be created from the hidden variable $\mathbf{x}$.

One can note the resemblance between Equation 2.34 and Equation 2.21. Both come from a constraint imposed in the probability distribution. There it was the expected value for the Hamiltonian over an uniform prior and now the constraint was the expected value of the generator functions over a known, but not constant, prior.

We also have the same difficulty than in statistical mechanics and thermodynamics of changing between intensive and extensive variables (here, λ 's and η 's, although they need not be intensive or extensive for the maximum entropy method to work). This change between representations can be a difficulty when performing the differentiation in Equation 2.34, but we surpass this problem in our social agent model by choosing a multivariate gaussian family.

2.5.2 Agent model definition

To continue with a model for an agent, we must first state the learning scenario over which the inference is being made. In our model the variable being inferred is a vector $\mathbf{x} = \mathbf{B} \in \mathbb{R}^K$.

The information available to perform the inference comes in the form of data $y = (\xi, \sigma)$ with $\xi \in \mathbb{R}^K$ and $\sigma \in \{\pm 1\}$. This represents a situation in which an agent is trying to learn how to classify an issue ξ with an opinion σ . This classification task induces a separation of $\mathbb{R}^K$ by a hyperplane and is consistent with the idea that the agents are evaluating whether they agree (+1) or disagree (-1) with the issue ξ .

We also hypothesize to have background information about a multiplicative noise ε with value between 0 and 1.

The likelihood distribution of the model, which is the distribution that describes the data generation process, is $P(\xi, \sigma | \mathbf{B}, \varepsilon) = P(\xi | \mathbf{B}, \varepsilon)P(\sigma | \xi, \mathbf{B}, \varepsilon)$. For simplicity we consider $P(\xi) = \delta(\xi - \mathcal{Z})$, an issue that is always being discussed in the society.

As mentioned before, the model for $P(\sigma | \xi, \mathbf{B}, \varepsilon)$ is that of binary linear classifier subject to multiplicative noise:

$$\begin{aligned} P(\sigma | \mathcal{Z}, \mathbf{B}, \varepsilon) &= \varepsilon \Theta(-\sigma \mathbf{B} \cdot \mathcal{Z}) + (1 - \varepsilon) \Theta(\sigma \mathbf{B} \cdot \mathcal{Z}) \\ &= \varepsilon + (1 - 2\varepsilon) \Theta(\sigma \mathbf{B} \cdot \mathcal{Z}) \end{aligned} \quad (2.35)$$

where Θ is the Heaviside step function.

Finally we must provide the generator functions of the model. We assume $\mathbb{E}_n[B^i] = J_n^i$ and $\mathbb{E}_n[B^i B^j] = C_n^{ij} + J_n^i J_n^j$ are the relevant generators, which yields a Multivariate Gaussian distribution family:

$$Q_n(\mathbf{B}) \equiv \mathcal{N}(\mathbf{B} | \mathbf{J}_n, \mathbf{C}_n) = |2\pi \mathbf{C}_n|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{B} - \mathbf{J}_n) \cdot \mathbf{C}_n^{-1} (\mathbf{B} - \mathbf{J}_n) \right] \quad (2.36)$$

$$= \frac{1}{Z_{\mathcal{N}_n}} \exp \left[-\vec{\lambda}_n \cdot \mathbf{B} - \mathbf{B} \cdot \underline{\lambda}_n \mathbf{B} \right] \quad (2.37)$$

where $\lambda_n^i = -(C_n^{-1})_{ij} J_n^j$ and $\underline{\lambda}_n^{ij} = \frac{1}{2} (C_n^{-1})_{ij}$ are the Lagrange multipliers that constrain the distribution.

One can think of this model as $\mathbf{J}$ being an *student* learning how to imitate a *professor* $\mathbf{B}$. It has information about the classifications the professor emits about several issues: $\mathcal{D}_n = \{(\xi_\mu, \sigma_\mu)_{\mu=1}^n\}$, which is also called learning set at time n . This is where we start to see our understanding of a **social agent**: when we couple many of these machines together they start learning from one another, being both students and professors at different timesteps.

Luckily we can invert the derivatives with respect to the Lagrange multipliers to a derivative with respect to the expected values of our generator functions⁹:

$$\frac{\partial}{\partial \lambda_n^i} = \frac{\partial J_n^l}{\partial \lambda_n^i} \frac{\partial}{\partial J_n^l} = -C_n^{li} \frac{\partial}{\partial J_n^l} \quad (2.38)$$

Now we can come back to Equation 2.34 and find

$$J_{n+1}^i = J_n^i - \frac{\partial \log Z_{n+1}}{\partial \lambda_n^i} = J_n^i - C_n^{li} \frac{\partial \log Z_{n+1}}{\partial J_n^l} \quad (2.39)$$

Noticing that $\mathbf{C}$ is symmetric because it is a covariance matrix, we can write in vectorial form:

$$\mathbf{J}_{n+1} = \mathbf{J}_n + \mathbf{C}_n \nabla_{\mathbf{J}_n} \log Z_{n+1} \quad (2.40)$$

We could follow the same procedure to study the evolution of C_n^{ij} but in that case we would need to study the derivative of $\log Z_{n+1}$ with respect to $(C_n^{-1})_{ij}$, which can be a complicated endeavour. We prefer to study the update through the analogous procedure of matching the generators' expected values:

$$\begin{aligned} \mathbb{E}_{Q_{n+1}} [B^i B^j] &\equiv C_{n+1}^{ij} + J_{n+1}^i J_{n+1}^j \\ &= \frac{1}{Z_{n+1}} \int d\mathbf{B} B^i B^j L_{n+1} Q_{n+1} \equiv \mathbb{E}_{P_{n+1}} [B^i B^j] \end{aligned} \quad (2.41)$$

In order to proceed with this calculation we must find an useful result:

$$\begin{aligned} \frac{\partial Q_n}{\partial J_n^k} &= Q_n \left[-\frac{1}{2} \sum_{i,j} (C_n^{-1})_{ij} \left((-2) B^i \frac{\partial J_n^j}{\partial J_n^k} + \frac{\partial}{\partial J_n^k} (J_n^i J_n^j) \right) \right] \\ &= Q_n \left[\sum_i (C_n^{-1})_{ik} (B^i - J_n^i) \right] \end{aligned} \quad (2.42)$$

⁹Remember Einstein summation convention: $\frac{\partial J_n^l}{\partial \lambda_n^i} \frac{\partial}{\partial J_n^l}$ implies a sum $\sum_{l=1}^K$

From that, we obtain

$$\nabla_{\mathbf{J}_n} Q_n = Q_n \mathbf{C}_n^{-1} (\mathbf{B} - \mathbf{J}_n) \quad (2.43)$$

$$\mathbf{C}_n \nabla_{\mathbf{J}_n} Q_n = Q_n (\mathbf{B} - \mathbf{J}_n) \quad (2.44)$$

$$B^i Q_n = J_n^i Q_n + \sum_k C_n^{ik} \frac{\partial Q_n}{\partial J_n^k} \quad (2.45)$$

Now we can go back to Equation 2.42 and use the result obtained in Equation 2.45 to get:

$$C_{n+1}^{ij} + J_{n+1}^i J_{n+1}^j = \frac{1}{Z_{n+1}} \int d\mathbf{B} L_{n+1} B_i \left(J_n^j Q_n + \sum_k C_n^{jk} \frac{\partial Q_n}{\partial J_n^k} \right) \quad (2.46)$$

$$= \frac{1}{Z_{n+1}} \left(J_n^j \int d\mathbf{B} L_{n+1} B_i Q_n + \sum_k C_n^{jk} \frac{\partial}{\partial J_n^k} \int d\mathbf{B} L_{n+1} B_i Q_n \right) \quad (2.47)$$

$$= \frac{1}{Z_{n+1}} \left[J_n^j \int d\mathbf{B} L_{n+1} \left(J_n^i Q_n + \sum_k C_n^{ik} \frac{\partial Q_n}{\partial J_n^k} \right) + \sum_k C_n^{jk} \frac{\partial}{\partial J_n^k} \int d\mathbf{B} L_{n+1} \left(J_n^i Q_n + \sum_l C_n^{il} \frac{\partial Q_n}{\partial J_n^l} \right) \right] \quad (2.48)$$

Applying the derivatives and substituting the integrals as the definition for Z , we obtain:

$$C_{n+1}^{ij} + J_{n+1}^i J_{n+1}^j = \frac{1}{Z_{n+1}} \left[J_n^i J_n^j Z_{n+1} + J_n^j \sum_k C_n^{ik} \frac{\partial}{\partial J_n^k} Z_{n+1} + J_n^i \sum_k C_n^{jk} \frac{\partial}{\partial J_n^k} Z_{n+1} + C_n^{ji} Z_{n+1} + \sum_k \sum_l C_n^{jk} C_n^{il} \frac{\partial}{\partial J_n^k} \frac{\partial}{\partial J_n^l} Z_{n+1} \right] \quad (2.49)$$

Now we cancel the Z_{n+1} terms that can be canceled and incorporate the others into the derivatives, using the fact that the derivative of $\log Z_n$ is $\frac{1}{Z_n}$:

$$C_{n+1}^{ij} + J_{n+1}^i J_{n+1}^j = C_n^{ji} + J_n^i J_n^j - J_n^j \sum_k C_n^{ik} \frac{\partial \log Z_{n+1}}{\partial J_n^k} - J_n^i \sum_k C_n^{jk} \frac{\partial \log Z_{n+1}}{\partial J_n^k} + \frac{1}{Z_{n+1}} \sum_k \sum_l C_n^{jk} C_n^{il} \frac{\partial}{\partial J_n^k} \frac{\partial}{\partial J_n^l} Z_{n+1} \quad (2.50)$$

In order to simplify this expression we can appeal to two different results. First of all, we

can go back to Equation 2.39 and apply it into the left-hand side of the equation above:

$$\begin{aligned}
J_{n+1}^i J_{n+1}^j &= \left(J_n^i - \sum_k C_n^{ki} \frac{\partial \log Z_{n+1}}{\partial J_n^k} \right) \left(J_n^j - \sum_k C_n^{kj} \frac{\partial \log Z_{n+1}}{\partial J_n^k} \right) \\
&= J_n^i J_n^j - J_n^i \sum_k C_n^{kj} \frac{\partial \log Z_{n+1}}{\partial J_n^k} - J_n^j \sum_k C_n^{ki} \frac{\partial \log Z_{n+1}}{\partial J_n^k} \\
&\quad + \sum_k \sum_l C_n^{li} C_n^{kj} \frac{\partial \log Z_{n+1}}{\partial J_n^k} \frac{\partial \log Z_{n+1}}{\partial J_n^l}
\end{aligned} \tag{2.51}$$

Substituting back:

$$C_{n+1}^{ij} = C_n^{ji} + \sum_k \sum_l C_n^{jk} C_n^{il} \left[\frac{1}{Z_{n+1}} \frac{\partial}{\partial J_n^k} \frac{\partial}{\partial J_n^l} - \frac{\partial \log Z_{n+1}}{\partial J_n^k} \frac{\partial \log Z_{n+1}}{\partial J_n^l} \right] \tag{2.52}$$

Now we can calculate and invoke a second result:

$$-\frac{\partial}{\partial J_n^k} \frac{\partial}{\partial J_n^l} \log Z_{n+1} = \frac{\partial}{\partial J_n^k} \frac{\partial}{\partial J_n^l} \log Z_{n+1} = \frac{\partial}{\partial J_n^k} \left(\frac{1}{Z_{n+1}} \frac{\partial}{\partial J_n^l} Z_{n+1} \right) \tag{2.53}$$

$$= \frac{1}{Z_{n+1}} \frac{\partial}{\partial J_n^k} \frac{\partial}{\partial J_n^l} Z_{n+1} - \frac{1}{Z_{n+1}^2} \frac{\partial}{\partial J_n^k} Z_{n+1} \frac{\partial}{\partial J_n^l} Z_{n+1} \tag{2.54}$$

$$= \frac{1}{Z_{n+1}} \frac{\partial^2 Z_{n+1}}{\partial J_n^k \partial J_n^l} - \left(\frac{\partial}{\partial J_n^k} \log Z_{n+1} \right) \left(\frac{\partial}{\partial J_n^l} \log Z_{n+1} \right) \tag{2.55}$$

And finally, we have that:

$$C_{n+1}^{ij} = C_n^{ji} - \sum_k \sum_l C_n^{jk} C_n^{il} \frac{\partial}{\partial J_n^k} \frac{\partial}{\partial J_n^l} \log Z_{n+1} \tag{2.56}$$

Again, in vectorial form:

$$\mathbf{C}_{n+1} = \mathbf{C}_n + \mathbf{C}_n \left(\mathbf{H}_{\mathbf{J}_n} \log Z_{n+1} \right) \mathbf{C}_n \tag{2.57}$$

where $\mathbf{H}_{\mathbf{J}_n} \log Z_{n+1}$ is a notation for the matrix of second derivatives of $\log Z_{n+1}$ with respect to the elements of $\mathbf{J}_n$

Finally, Equations 2.40, 2.57 describe the update for the expected values $\mathbf{J}_n$ and $\mathbf{C}_n$, respectively. Now we only need to calculate the Z_{n+1} term to complete our entropic dynamics inference describing social agents.

2.5.3 Calculating Z_{n+1}

At last we can calculate the evidence of the model Z_{n+1}

$$Z_{n+1} = \int d\mathbf{x} P(\mathcal{D}_{n+1}|\mathbf{x})P(\mathbf{x}|\theta_n) = \int d\mathbf{B} P(\xi)P(\sigma|\xi, \mathbf{B})Q_n(\mathbf{B}) \quad (2.58)$$

$$= P(\xi) \langle P(\sigma|\xi, \mathbf{B}) \rangle_{Q_n(\mathbf{B})} \quad (2.59)$$

where $P(\xi)$ is independent of $\mathbf{B}$ so we can take it off the integral; it will not contribute when we differentiate $\log Z_{n+1}$.

Remembering the likelihood distribution from Equation 2.35, we now proceed to calculate the expected value of the Heaviside function:

$$\langle \Theta(\sigma \xi \cdot \mathbf{B}) \rangle_{Q_n(\mathbf{B})} = \int d\mathbf{B} \Theta(\sigma \xi \cdot \mathbf{B}) \frac{1}{|2\pi \mathbf{C}_n|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{B} - \mathbf{J}_n) \cdot \mathbf{C}_n^{-1} (\mathbf{B} - \mathbf{J}_n) \right] \quad (2.60)$$

$$= \int db \Theta(\sigma b) \int \frac{d\mathbf{B}}{|2\pi \mathbf{C}_n|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{B} - \mathbf{J}_n) \cdot \mathbf{C}_n^{-1} (\mathbf{B} - \mathbf{J}_n) \right] \delta \left(b - \frac{1}{\sqrt{K}} \xi \cdot \mathbf{B} \right) \quad (2.61)$$

$$= \int db \Theta(\sigma b) \int \frac{d\hat{b}}{2\pi} e^{i\hat{b}b} \int \frac{d\mathbf{B}}{|2\pi \mathbf{C}_n|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{B} - \mathbf{J}_n) \cdot \mathbf{C}_n^{-1} (\mathbf{B} - \mathbf{J}_n) - \frac{i\hat{b}}{\sqrt{K}} \xi \cdot \mathbf{B} \right] \quad (2.62)$$

$$= \int db \Theta(\sigma b) \int \frac{d\hat{b}}{2\pi} e^{i\hat{b}b} \exp \left[-\frac{1}{2} \left(\frac{\hat{b}^2}{K} \xi \cdot \mathbf{C}_n \xi + \frac{2i\hat{b}}{\sqrt{K}} \xi \cdot \mathbf{J}_n \right) \right] \quad (2.63)$$

$$= \int db \Theta(\sigma b) \int \frac{d\hat{b}}{2\pi} \exp \left[-\frac{1}{2} (\hat{b}^2 \Gamma_n^2 + 2i\hat{b}(h_n - b)) \right] \quad (2.64)$$

$$= \int db \Theta(\sigma b) \frac{1}{\sqrt{2\pi \Gamma_n^2}} \exp \left[-\frac{1}{2} \left(\frac{h_n - b}{\Gamma_n} \right)^2 \right] \quad (2.65)$$

where we defined $h_n = \frac{1}{\sqrt{K}} \mathcal{Z} \cdot \mathbf{J}_n$ and defined $\Gamma_n^2 = \frac{1}{K} \mathcal{Z} \cdot \mathbf{C}_n \mathcal{Z}$.

Analysing each case $\sigma = \pm 1$ separately, we end up with:

$$Z_{n+1} = P(\xi) \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{\sigma h_n}{\Gamma_n} \right) \right] \quad (2.66)$$

where Φ is the cumulative distribution function of the gaussian distribution.

Finally, taking the logarithm (and discarding the constant part $P(\xi)$ which will not contribute to our inference when we take derivatives):

$$-\log Z_{n+1} = -\log \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{\sigma h_n}{\Gamma_n} \right) \right] \quad (2.67)$$

This quantity is highly important because of Equations 2.34, 2.40, 2.57. It functions as a cost function, generating the dynamics of the model. Although it is not an energy, it will have the same role a Hamiltonian has in statistical mechanical systems (see Section 2.4.2): it is relevant and sufficient information to describe the evolution of our system.

3

Models

Entropic dynamics results shown in Section 2.5 can give rise to several different models and subsequent predictions, some of which we present in this chapter. Our main goal is to understand more about how agents interact in society, with an emphasis in: understanding group formation; assessing the possibility of categorizing and describing different political groups; and gauging the importance of selecting the issues to be discussed in a society.

The first model we present (Section 3.1) is a statistical mechanical one defined by the potential described in Equation 2.67. After a mean field approximation we obtain, in Section 4.1, a phase transition between an ordered and a disordered phase. This result was already known by (Caticha and Vicente 2011; Vicente et al. 2014) using similar models, and we show a different approach to obtain the expected outcome.

This model yields some interesting results regarding the phase transition between ordered and disordered societies, but it has two major drawbacks:

1. It is too intractable to perform comparisons with available data;
2. It cannot account for the existence of opposite and polarized groups without resorting to additional hypothesis.

Henceforth, we analyse alternative models in Section 3.2 and Section 3.3 that maintain some resemblance to the model before while also accounting for those 2 points.

The second model is based upon a simpler one studied by (Vicente et al. 2014). We compare predictions of this model with empirical data from the MFQ (Graham, Haidt, and Nosek 2009; Graham et al. 2013; Haidt 2007). We present the assumptions made to connect

data and model, and the inference steps needed to estimate from the questionnaire some parameters of the model. Also with this model, in Section 4.2, we try to characterize different political groups, such as conservatives and liberals, and we introduce the concept of extremists.

Finally, the third model extends the previous one to a setting in which we have two opposing parties that have an antiferromagnetic interaction between them, but ferromagnetic inside each group. This model presents an interesting interplay between having a cohesive “friends” group and opposing the “foes” group. This model touches the idea of categorization that (Abrams et al. 1990) mentioned: an agent tries to learn from peers of the same group and to anti-learn from peers of an opposite community.

3.1 Model 1

In our first model we are interested in studying the distribution of opinions in a society discussing one issue, the $\mathcal{Z}$. A possible strategy is to take a society of the agents developed in Section 2.5, recognize relevant information describing the model and proceed with Statistical Mechanics calculations. At some point the calculation might become intractable and one must transition to approximate results and/or computational methods. In this section we develop a Mean Field approach to an specific canonical ensemble of social agents in a noisy society. The work follows closely (Simões and Caticha 2018).

To look for the relevant information that describes the model we first look at the Update Equations 2.40, 2.57. For simplicity we consider that the description of our “moral space” $\mathbb{R}^K$ is already one that renders the “moral dimensions” independent from one another, and we assume that $\mathbf{C}_n = \gamma_n^2 \mathbb{1}$. We also assume that, for a certain timescale, the evolution of γ is frozen. Hence, the update mechanism is going to be led only by:

$$\mathbf{J}_{n+1} = \mathbf{J}_n - \nabla_{\mathbf{J}_n} \gamma^2 \log \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{\sigma h_n}{\gamma} \right) \right] \quad (3.1)$$

Under those circumstances, we assume the term inside the differentiation is relevant and sufficient information to describe the evolution of our system, and we can consider that our society of agents $\{J^i\}$ can be described totally by one specific Hamiltonian $\mathcal{H}$:

$$\mathcal{H} = -\gamma^2 \sum_{\langle i,j \rangle} \log \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{1}{\gamma} (\sigma_i h_j + \sigma_j h_i) \right) \right] = \sum_{\langle i,j \rangle} V_{ij} \quad (3.2)$$

We can study this society using entropic inference, which lead us to the following canonical (Boltzmann) probability distribution:

$$P_B(\{\mathbf{J}_i\}) = \frac{1}{Z_B} \exp(-\beta \mathcal{H}(\{\mathbf{J}_i\})) \quad (3.3)$$

This description, however, is rather complex depending on the kind of applications we want to follow. A common procedure to simplify our model is to consider a mean field approximation, which projects our solution (the probability distribution P_B , which depends on the whole set $\{\mathbf{J}_i\}$ all at once) into a parametric family of separable probability distributions $P_0 = \prod_i P_i(\mathbf{J}_i)$ much simpler to work with.

In that case, we do not wish to choose a separable distribution indiscriminately; we want to pick as close as possible the specific P_0 which best approximates P_B given the constraints we have assigned to it. That is a calculation we can do maximizing the entropy S (or minimizing the Kullback-Leibler divergence), as follows:

$$S[P_0||P_B] = - \iint \left(\prod_{i=1}^N d\mathbf{J}_i \right) P_0 \log \left(\frac{P_0}{P_B} \right) \quad (3.4)$$

$$= - \iint \left(\prod_i d\mathbf{J}_i \right) P_0 [\log(Z_B) + \beta \mathcal{H} + \log(\prod_i P_i)] \quad (3.5)$$

$$= \left\langle \log Z_B - \beta \mathcal{H} - \sum_i \log P_i \right\rangle_{P_0} \quad (3.6)$$

We want to find P_0 so that variations of S equals to zero, $\delta S = 0$. Usual functional calculus arguments lead to:

$$\delta S = \int dJ_j \delta P_j \left[\log Z_B + \log P_j + 1 + \beta \sum_{i \in \partial j} \int d\mathbf{J}_i P_i V_{ij} \right] \quad (3.7)$$

The desired result can only be obtained for any variation δP_j if the term in brackets [...] is identically zero itself. We get the following:

$$P_j(\mathbf{J}_j) = \frac{1}{Z_j} \exp \left(-\beta \sum_{i \in \partial j} \int d\mathbf{J}_i P_i V_{ij} \right) \quad (3.8)$$

where Z_j is the new partition function/normalization of the model.

Inserting back V_{ij} (Equation 3.2):

$$P_j(\mathbf{J}_j) = \frac{1}{Z_j} \exp \left(\beta \gamma^2 \sum_{i \in \partial j} \int d\mathbf{J}_i P_i \log \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{\sigma_i h_j + \sigma_j h_i}{\gamma} \right) \right] \right) \quad (3.9)$$

Unfortunately, due to the rather complex form of the potential V_{ij} , the integral in Equation 3.9 is intractable. In that case, instead of selecting the best mean field probability distribution family, we are going to choose an approximation with a similar functional form.

Let us consider the fact that in a mean field model agent j does not interact directly with another agent i , but only with an external effective field. With that in mind we can approximate the integral over i in Equation 3.9 to $\log \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{r h_j + \sigma_j m}{\gamma} \right) \right]$. The sum over neighbors ∂j becomes an effective number of neighbors for agent j that can also be approximated to a constant value ν throughout society. Although people have different number of peers depending on their sociability, we expect them to have similar values nonetheless.

Taking the logarithm out of the exponential, one obtains:

$$P_j(\mathbf{J}_j) = \frac{1}{Z_j} \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{1}{\gamma} (r h_j + m \sigma_j) \right) \right]^{\beta \nu \gamma^2} \quad (3.10)$$

Now one has to interpret the meaning of parameters m and r to proceed. One can see them as parameters that describe the overall behavior of the society, such that one agent i receives signals from its neighbors irrespective of label i . In fact, remembering they came from the integral in Equation 3.9, we can see them as expected values over the distribution $P_{\text{MF}}(\mathbf{J})$

We can represent this self-consistently with the following set of equations to be obeyed:

$$m = \int d\mathbf{J} h(\mathbf{J}) P_{\text{MF}}(\mathbf{J}) \quad (3.11)$$

$$r = \int d\mathbf{J} \text{sign } h(\mathbf{J}) P_{\text{MF}}(\mathbf{J}) \quad (3.12)$$

And our final mean field probability distribution becomes:

$$P_{\text{MF}}(\mathbf{J}) = \frac{1}{Z} \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{r}{\gamma} h(\mathbf{J}) + \frac{m}{\gamma} \text{sign } h(\mathbf{J}) \right) \right]^{\beta \nu \gamma^2} \quad (3.13)$$

This result does not mean that all the agents are identical, but that they draw their moral vector from the same probability distribution (being held constant parameters $\beta, \varepsilon, \gamma$).

We can also rewrite our order parameters

$$\begin{aligned}
 m &= \int d\mathbf{J}_i h_i P_i = \frac{1}{Z_i} \int d\mathbf{J}_i h_i \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{1}{\gamma} (r h_i + \sigma_i m) \right) \right]^{\beta \nu \gamma^2} \\
 &= \frac{1}{Z_i} \frac{1}{\sqrt{K}} \int d\mathbf{J}_i \mathcal{Z} \cdot \mathbf{J}_i \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{1}{\gamma} \left(r \frac{\mathcal{Z} \cdot \mathbf{J}_i}{\sqrt{K}} + \text{sign}(\mathcal{Z} \cdot \mathbf{J}_i) m \right) \right) \right]^{\beta \nu \gamma^2}
 \end{aligned} \tag{3.14}$$

$$\begin{aligned}
 r &= \int d\mathbf{J}_i \sigma_i P_i = \frac{1}{Z_i} \int d\mathbf{J}_i \sigma_i \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{1}{\gamma} (r h_i + \sigma_i m) \right) \right]^{\beta \nu \gamma^2} \\
 &= \frac{1}{Z_i} \int d\mathbf{J}_i \text{sign}(\mathcal{Z} \cdot \mathbf{J}_i) \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{1}{\gamma} \left(r \frac{\mathcal{Z} \cdot \mathbf{J}_i}{\sqrt{K}} + \text{sign}(\mathcal{Z} \cdot \mathbf{J}_i) m \right) \right) \right]^{\beta \nu \gamma^2}
 \end{aligned} \tag{3.15}$$

Since we can always rotate the coordinate system to a given orientation, we choose one in which $\mathcal{Z} = |\mathcal{Z}| \hat{\mathbf{e}}_5$, so that $\mathcal{Z} \cdot \mathbf{J}_i = \sqrt{K} \cos \theta$ and all the other angular integrals are trivial besides the one in θ :

$$\begin{aligned}
 m &= \frac{1}{\zeta} \int_0^\pi d\theta \sin^3 \theta \cos \theta B(\theta | \varepsilon, \gamma, m, r, \beta \nu) \\
 r &= \frac{1}{\zeta} \int_0^\pi d\theta \sin^3 \theta \text{sign}(\cos \theta) B(\theta | \varepsilon, \gamma, m, r, \beta \nu) \\
 \zeta &= \int_0^\pi d\theta \sin^3 \theta B(\theta | \varepsilon, \gamma, m, r, \beta \nu)
 \end{aligned} \tag{3.16}$$

where we implicitly defined the function $B(\theta) \equiv B(\theta | \varepsilon, \gamma, m, r, \beta \nu)$:

$$B(\theta) = \left[\varepsilon + (1 - 2\varepsilon) \Phi \left(\frac{1}{\gamma} (r \cos \theta + \text{sign}(\cos \theta) m) \right) \right]^{\beta \nu \gamma^2} \tag{3.17}$$

We now proceed to the numerical determination of values of (m, r, ζ) for each set of parameters $(\varepsilon, \gamma, \beta \nu)$ by integrating Equation 3.16. We present this in Section 4.1.

3.2 Model 2

As mentioned at the beginning of Section 3, the model presented in Section 3.1, although presenting some interesting features, is not suitable to make comparisons with data. In this second model we look at the same problem of characterizing the distribution of opinions in a society, but now with a different approach. We build upon the development presented in (Vicente et al. 2014), where following interaction Hamiltonian in a society was introduced:

$$\mathcal{H} = -\frac{1+\delta}{2} \sum_{\langle ij \rangle} h_i h_j + \frac{1-\delta}{2} \sum_{\langle ij \rangle} |h_i h_j| \quad (3.18)$$

where δ is a “psychological cost” of agreement between agents, with $0 < \delta < 1$. When $\delta = 0$ the agents are called “error correctors”, because they do not change their moral vectors in the case of matching opinions, and when $\delta = 1$ they are called “corroboration seekers”, because they attribute the same amount of information to both cases: agreement or disagreement.

On this same paper, the authors develop a mean-field approximation using Equation 3.18 and the Maximum Entropy method, similar to the one performed in Section 3.1 to arrive at Equation 3.8. The information constraint on the expected value of the energy $\langle \mathcal{H} \rangle = E$ enforces the appearance of a Lagrange multiplier β , the “social pressure” of the society. The resulting distribution is:

$$P_{\text{MF}}(h) = \frac{\tilde{\beta}^2}{2} (1 - h^2) e^{-\tilde{\beta}(1-h)} \quad (3.19)$$

where $\tilde{\beta}$ is a parameter aggregating β and δ . As before, h is the dot product between the agent’s inner moral representation $\mathbf{J}$ and a discussed query (Zeitgeist) $\mathcal{Z}$.

We should also note there is an assumption that $h > 0$ made to get to Equation 3.19. This enforces a cohesive society, because the magnetization $\langle h \rangle$ will be positive as well. To some extent we break this hypothesis later in Section 3.3.

Assuming one has access to responses from several people of one same group g . This group is a collective of people that shares similar cultural and moral values. We make the hypothesis that all responses obtained from subjects of that were sampled from the same distribution $P(h|\tilde{\beta}_g)$.

In our case, we are going to use the different political affiliations (pa) the respondents of the MFQ attribute to themselves. There are also other relevant information on the dataset that we chose to ignore on this first analysis, for example information on the country of the respondent, or the age. The distribution we sample from, therefore, is $P(h|\tilde{\beta}_{pa})$ for a given

political affiliation pa .

This choice of groups is not the same as saying that all liberals are identical, or that a liberal agent cannot have the same opinion as a conservative one on a specific topic; we are only saying that there must be some similarity between people that think alike, they ought to be characterized in groups, and we can describe this using a probability distribution.

Then, one can use Equation 3.19 and apply Bayes' theorem to it to find an estimator for $\tilde{\beta}$ given a set of data $\{h_i\}_{i \in pa}$:

$$P(\tilde{\beta}|\{h_i\}_{i \in pa}) \propto P(\tilde{\beta}) \prod_{i \in pa} P(h_i|\tilde{\beta}) \quad (3.20)$$

We note that $P(h|\tilde{\beta})$ resembles a Gamma distribution in $\tilde{\beta}$ and choose a prior distribution $\tilde{\beta} \sim \text{Gamma}(k_0, \theta_0)$, that is:

$$P(\tilde{\beta}) = \frac{\theta_0^{-k_0}}{\Gamma(k_0)} \tilde{\beta}^{k_0-1} e^{-\tilde{\beta}/\theta_0} \quad (3.21)$$

This yields a Gamma distribution as a posterior as well. First we evaluate the normalization factor $P(\{h_i\}_{pa})$:

$$P(\{h_i\}) = \int_0^\infty d\tilde{\beta} P(\{h_i\}, \tilde{\beta}) \quad (3.22)$$

$$= \int_0^\infty d\tilde{\beta} \frac{\theta_0^{-k_0}}{\Gamma(k_0)} \tilde{\beta}^{k_0-1} e^{-\tilde{\beta}/\theta_0} \prod_i \frac{\tilde{\beta}^2}{2} (1 - h_i^2) e^{-\tilde{\beta}(1-h_i)} \quad (3.23)$$

$$= \int_0^\infty d\tilde{\beta} \frac{\theta_0^{-k_0}}{\Gamma(k_0)} \tilde{\beta}^{k_0-1} e^{-\tilde{\beta}/\theta_0} \frac{\tilde{\beta}^{2n}}{2^n} e^{-\tilde{\beta}(n-m)} \prod_i (1 - h_i^2) \quad (3.24)$$

$$= \frac{\theta_0^{-k_0}}{2^n \Gamma(k_0)} \prod_i (1 - h_i^2) \int_0^\infty d\tilde{\beta} \tilde{\beta}^{2n+k_0-1} e^{-\tilde{\beta}(n+1/\theta_0-m)} \quad (3.25)$$

where n is the number of respondents in the group being considered and $m = \sum_i h_i$ is analogous to a magnetization of the group.

Recognize the Gamma integral $\Gamma(z) = \int_0^\infty dt t^{z-1} e^{-t}$ with $z = 2n+k_0$ to do the integration. Making the substitution $t = \tilde{\beta}(n - m + 1/\theta_0)$, we obtain:

$$P(\{h_i\}) = \frac{\theta_0^{-k_0}}{2^n \Gamma(k_0)} (n + 1/\theta_0 - m)^{-2n-k_0+1} \prod_i (1 - h_i^2) \int_0^\infty dt t^{2n+k_0-1} e^{-t} \quad (3.26)$$

$$= \frac{1}{2^n} \frac{\theta_0^{-k_0}}{\left(n + \frac{1}{\theta_0} - m\right)^{2n+k_0-1}} \frac{\Gamma(2n + k_0)}{\Gamma(k_0)} \prod_i (1 - h_i^2) \quad (3.27)$$

Inserting back into Bayes' rule, we obtain the posterior $\tilde{\beta}|\{h_i\}_{i \in pa} \sim \text{Gamma}\left(k_0 + 2n, \frac{\theta_0}{1+\theta_0(n-m)}\right)$. Explicitly:

$$P(\tilde{\beta}|\{h_i\}_{i \in pa}) = \frac{\left(\frac{1}{\theta_0} + n_{pa} - m_{pa}\right)^{k_0+2n_{pa}}}{\Gamma(k_0 + 2n_{pa})} \tilde{\beta}^{k_0-1+2n_{pa}} e^{-\tilde{\beta}\left(\frac{1}{\theta_0} + n_{pa} - m_{pa}\right)} \quad (3.28)$$

This posterior is used to predict $\tilde{\beta}$ for a given dataset. The advantage of using a posterior distribution instead of a point estimate is that we have more information about the inference being made. In a case where one finds a bimodal posterior distribution, for example, one knows that a mean value estimate would not describe well the data. In the specific case of Section 4.2 we use it on MFQ data for different political affiliations, and luckily we obtained as a sharp distribution, which justifies a maximum a posteriori or a mean value estimates.

3.3 Model 3

Both models presented before have the disadvantage of not being able to describe situations in which there is a subdivision of factions, or communities, inside a same society. Other results from our group (Alves and Caticha 2016, 2018) give us reason to believe that the multiplicative noise ε is an important parameter to analyze this phenomena of group formation and opposition between different groups.

This has led us into developing a third model, which would capture the intuitions we obtained with the first two models and could treat this situation of opposing communities' interactions. Our first model, despite being a cumbersome model to work with, had the parameter ε naturally imbued into it. However, this model only yields reasonable results for $\varepsilon < 0.5$. This is because for values of $\varepsilon > 0.5$ the society enters a frustrated state which cannot be resolved by aligning nor anti-aligning the opinions of the agents. Our second model, while being more amenable to calculations, did not have a parameter enabling the existence of antiferromagnetic interactions.

Our solution then, is to develop a model from the second model (because of ease of analytical manipulation and of interpretation of results) while still capturing the existence of

a parameter that describes opposition and antiferromagnetism, ε .

We build upon ideas of the infinite-range Ising model with antiferromagnetic interactions, which we present below.

3.3.1 Infinite-range Ising model

The hamiltonian describing the infinite-range Ising model is

$$\mathcal{H}_I = -\frac{J}{N} \sum_{\langle ij \rangle} s_i s_j - h \sum_i s_i \quad (3.29)$$

Considering the constraint of the expected value of the energy $\langle \mathcal{H} \rangle = E$, usual maximum entropy arguments lead us into a canonical distribution $P_I(\{s_i\}) = \frac{1}{Z_I} \exp(-\beta \mathcal{H})$. We proceed evaluating the partition function of the model.

$$Z_I = \sum_{\{s\}} \exp \left(\frac{\beta J}{N} \sum_{\langle ij \rangle} s_i s_j + \beta h \sum_i s_i \right) \quad (3.30)$$

We cannot do the sum to solve Z_I with the quadratic terms ($s_i s_j$) in the exponential, so we use the following expression taken from the normalization of a gaussian distribution:

$$\exp \left(\frac{\beta J N \mu^2}{2} \right) = \int_{-\infty}^{\infty} \frac{dM}{\sqrt{2\pi/(\beta J N)}} \exp \left(-\frac{1}{2} \beta J N M^2 + \beta J N \mu M \right) \quad (3.31)$$

Now we define $\mu = \frac{1}{N} \sum_i s_i$ and substitute Equation 3.31 into Equation 3.30 to perform the sum and obtain:

$$Z_I = 2^N \int_{-\infty}^{\infty} \frac{dM}{\sqrt{2\pi/(\beta J N)}} \exp \left(-\frac{1}{2} \beta J N M^2 \right) \exp (N \log \cosh(\beta J M + \beta h)) \quad (3.32)$$

$$= c_N \int_{-\infty}^{\infty} dM e^{\beta J N f(M)} \quad (3.33)$$

where c_N is a collection of the constants and $f(M) = -\frac{1}{2} M^2 + \frac{1}{\beta J} \log \cosh(\beta J M + \beta h)$ can be associated to the free energy of the model.

We expect N to be large, so we expand $f(M)$ around its maximum M^* to obtain a good approximation of the integral. This is known as Laplace's method.

Differentiating f with respect to M and finding the maximum, one obtains the known relation of Curie-Weiss:

$$M^* = \tanh[\beta J M^* + \beta h] \quad (3.34)$$

Back to the partition function in Equation 3.33, we approximate the integral and obtain the free energy of the model, $-f(M^*)$:

$$\frac{1}{\beta J N} \log Z_I = f(M^*) + \text{constant terms} \quad (3.35)$$

One caveat of this calculation is that, for it to work properly, it depends on the fact that $J > 0$. This is because, when the external field h is zero, if we force $J < 0$ at Equation 3.34, the only solution becomes a zero magnetization system, the same as mentioned about our the first model with $\varepsilon > 0.5$.

This motivates the separation of the Ising spins into 2 groups: inside each group the interaction is ferromagnetic and between spins of different groups it is antiferromagnetic. The new hamiltonian then is:

$$\mathcal{H}_{AI} = -\frac{J_+}{N} \sum_{\langle ij \rangle \in A} s_i s_j - \frac{J_-}{N} \sum_{\langle ij \rangle \in B} s_i s_j + \frac{J_-}{N} \sum_{\substack{i \in A \\ j \in B}} s_i s_j - H_A \sum_{i \in A} s_i - H_B \sum_{j \in B} s_j \quad (3.36)$$

where $J_+, J_- > 0$.

Using the definitions $\mu_A = \frac{1}{N} \sum_{i \in A} s_i$ and $\mu_B = \frac{1}{N} \sum_{j \in B} s_j$, we can rewrite it as:

$$\mathcal{H}_{AI} = -\frac{N}{2} [J_+ \mu_A^2 + J_- \mu_B^2 - 2J_- \mu_A \mu_B] - N H_A \mu_A - N H_B \mu_B \quad (3.37)$$

We recognize that once more the squared terms in the Hamiltonian render the evaluation of the partition function difficult. We employ the following identity from the normalization of a multivariate gaussian distribution:

$$\exp\left(\frac{1}{2} \vec{\mu} \cdot \Sigma^{-1} \vec{\mu}\right) = \int \frac{d\mathbf{x}}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2} \mathbf{x} \cdot \Sigma^{-1} \mathbf{x}\right) \exp(\mathbf{x} \cdot \Sigma^{-1} \vec{\mu}) \quad (3.38)$$

where $|\Sigma|$ means the determinant of matrix Σ .

Matching

$$\Sigma^{-1} = \beta N \begin{bmatrix} J_+ & -J_- \\ -J_- & J_+ \end{bmatrix} \quad \vec{\mu} = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad (3.39)$$

we rewrite the partition function and perform the sum over spins to obtain:

$$Z = \frac{\beta N \sqrt{J_+^2 - J_-^2}}{2\pi} \int dx dy e^{\beta N f(x,y)} \quad (3.40)$$

$$\begin{aligned} f(x, y) = & -\frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} \begin{bmatrix} J_+ & -J_- \\ -J_- & J_+ \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \frac{N_A}{\beta N} \log \cosh [\beta (H_A + J_+ x - J_- y)] \\ & + \frac{N_B}{\beta N} \log \cosh [\beta (H_B + J_+ y - J_- x)] \end{aligned} \quad (3.41)$$

Differentiating and finding the maxima $f(x^*, y^*)$:

$$\begin{aligned} J_+ x^* - J_- y^* &= \frac{J_+ N_A}{N} \tanh [\beta (H_A + J_+ x^* - J_- y^*)] - \frac{J_- N_B}{N} \tanh [\beta (H_B + J_+ y^* - J_- x^*)] \\ J_+ y^* - J_- x^* &= \frac{J_+ N_B}{N} \tanh [\beta (H_B + J_+ y^* - J_- x^*)] - \frac{J_- N_A}{N} \tanh [\beta (H_A + J_+ x^* - J_- y^*)] \end{aligned}$$

After some steps, those are the same as the ones below, for null external fields H_A, H_B :

$$x^* = \frac{N_A}{N} \tanh (\beta J_+ x^* - \beta J_- y^*) \quad (3.42)$$

$$y^* = \frac{N_B}{N} \tanh (\beta J_+ y^* - \beta J_- x^*) \quad (3.43)$$

which now yields non-trivial results for (x^*, y^*) when $J_- > J_+$, the situation in which the antiferromagnetic interaction is stronger than the ferromagnetic one.

3.3.2 Bipartite Society

Both because of our interest in investigating groups' interactions in society and because of the drawback of the first model for $\varepsilon > 0.5$ just mentioned, we proceed in the same way of the infinite-range Ising model: we separate the agents into 2 groups, which we call *communities*. These communities have ferromagnetic interactions within them and antiferromagnetic ones between elements of different groups (depicted in Figure 3.1).

Figure 3.1: Exemplification of the interactions between and inside the 2 communities in a bipartite society

We borrow Equation 3.18 from Section 3.2, and consider two different situations to build the Hamiltonian of a bipartite society: the first one, in which the agents interacting are agreeing $h_i h_j > 0$; and a second one, in which the agents are disagreeing $h_i h_j < 0$. These two situations yield different contributions: $-\delta h_i h_j$ and $-h_i h_j$, respectively. The first term will be the interaction term inside a given community, whereas the second one will be the interaction term between communities.

After modulating the antiferromagnetic interaction with a noise level ε , the Hamiltonian describing our bipartite society will become:

$$\mathcal{H} = -\frac{\delta}{N} \sum_{\langle i,j \rangle \in A} h_i h_j - \frac{\delta}{N} \sum_{\langle i,j \rangle \in B} h_i h_j - \frac{(1-2\varepsilon)}{N} \sum_{\substack{i \in A \\ j \in B}} h_i h_j \quad (3.44)$$

where the subscript b stands for bipartite and the $h_i \in [-1, 1]$ are the opinion fields for each of the N agents. Both parameters δ and ε can have values on the interval $[-1, 1]$. We note that δ describes the intensity of the in-group ferromagnetic interactions, whereas ε describes the inter-group antiferromagnetic ones.

First, let us rewrite the Hamiltonian using the definitions $\mu_A = \frac{1}{N} \sum_{i \in A} h_i$ and $\mu_B = \frac{1}{N} \sum_{j \in B} h_j$:

$$\begin{aligned} \mathcal{H}_b &= -\frac{N}{2} \left[\delta \left(\frac{1}{N} \sum_{i \in A} h_i \right)^2 + \delta \left(\frac{1}{N} \sum_{j \in B} h_j \right)^2 + 2(1-2\varepsilon) \left(\frac{1}{N} \sum_{i \in A} h_i \right) \left(\frac{1}{N} \sum_{j \in B} h_j \right) \right] \\ &= -\frac{N}{2} [\delta \mu_A^2 + \delta \mu_B^2 + 2(1-2\varepsilon) \mu_A \mu_B] \end{aligned} \quad (3.45)$$

Once again we have squared terms in the Hamiltonian, which hinders the evaluation of Z_b . We use the gaussian identity from Equation 3.38 to simplify our expression for Z_b .

The terms being compared are:

$$\mu_A = \frac{1}{N} \sum_{i \in A} h_i \quad \mu_B = \frac{1}{N} \sum_{j \in B} h_j \quad (3.46)$$

$$\Sigma^{-1} = \beta N \begin{bmatrix} \delta & 1 - 2\varepsilon \\ 1 - 2\varepsilon & \delta \end{bmatrix} \quad (3.47)$$

Then, making the substitution above, the partition function becomes:

$$Z_b = \int_{-1}^1 dh_1 \int_{-1}^1 dh_2 \cdots \int_{-1}^1 dh_N \exp(-\beta \mathcal{H}) = \int dh \exp(-\beta \mathcal{H}) \quad (3.48)$$

$$= \int dh \exp \left\{ \frac{\beta N}{2} [\delta \mu_A^2 + \delta \mu_B^2 + 2(1 - 2\varepsilon) \mu_A \mu_B] \right\} \quad (3.49)$$

$$= \int dh \int_{R^2} \frac{dx dy}{\sqrt{|2\pi\Sigma|}} \exp \left(-\frac{1}{2} \mathbf{x} \cdot \Sigma^{-1} \mathbf{x} + \mathbf{x} \cdot \Sigma^{-1} \vec{\mu} \right) \quad (3.50)$$

$$= \int_{R^2} \frac{dx dy}{\sqrt{|2\pi\Sigma|}} \exp \left(-\frac{1}{2} \mathbf{x} \cdot \Sigma^{-1} \mathbf{x} \right) \int dh \exp(\mathbf{x} \cdot \Sigma^{-1} \vec{\mu}) \quad (3.51)$$

$$= \int_{R^2} \frac{dx dy}{\sqrt{|2\pi\Sigma|}} \exp \left(-\frac{\beta N}{2} (\delta x^2 + \delta y^2 + 2(1 - 2\varepsilon)xy) \right) \\ \times \int dh \exp(\beta N \mu_A [\delta x + (1 - 2\varepsilon)y] + \beta N \mu_B [\delta y + (1 - 2\varepsilon)x]) \quad (3.52)$$

Now we evaluate the integral over h separating into contributions from each group:

$$I_h = \int dh_A \exp \left(\beta [\delta x + (1 - 2\varepsilon)y] \sum_{i \in A} h_i \right) \int dh_B \exp \left(\beta [\delta y + (1 - 2\varepsilon)x] \sum_{j \in B} h_j \right) \quad (3.53)$$

$$= \left[\int_{-1}^1 dh_i \exp(\beta [\delta x + (1 - 2\varepsilon)y] h_i) \right]^{N_A} \left[\int_{-1}^1 dh_j \exp(\beta [\delta y + (1 - 2\varepsilon)x] h_j) \right]^{N_B} \quad (3.54)$$

$$= 2^N \left[\frac{\sinh \beta [\delta x + (1 - 2\varepsilon)y]}{\beta [\delta x + (1 - 2\varepsilon)y]} \right]^{N_A} \left[\frac{\sinh \beta [\delta y + (1 - 2\varepsilon)x]}{\beta [\delta y + (1 - 2\varepsilon)x]} \right]^{N_B} \quad (3.55)$$

Inserting this result back into Z , one obtains:

$$Z_b = 2^N \int_{R^2} \frac{dx dy}{2\pi} \sqrt{\beta \delta N} e^{Nf(x,y)} \quad (3.56)$$

where f will be associated to the free energy of the model:

$$f(x, y) = \frac{N_A}{N} \log \left(\frac{\sinh \beta [\delta x + (1 - 2\varepsilon)y]}{\beta [\delta x + (1 - 2\varepsilon)y]} \right) + \frac{N_B}{N} \log \left(\frac{\sinh \beta [\delta y + (1 - 2\varepsilon)x]}{\beta [\delta y + (1 - 2\varepsilon)x]} \right) \quad (3.57)$$

$$- \frac{\beta \delta}{2} (x^2 + y^2) - \beta(1 - 2\varepsilon)xy$$

Once again, we employ Laplace's method for approximation an integral by the maximum of its argument. Differentiating f with respect to its arguments, x and y , yields the following set of consistent equations to be satisfied at the maximum:

$$N [\delta x^* + (1 - 2\varepsilon)y^*] = N_A \delta \left[\coth \beta \delta x^* + (1 - 2\varepsilon)y^* - \frac{1}{\beta \delta x^* + (1 - 2\varepsilon)y^*} \right] \quad (3.58)$$

$$+ N_B (1 - 2\varepsilon) \left[\coth \beta \delta y^* + (1 - 2\varepsilon)x^* - \frac{1}{\beta \delta y^* + (1 - 2\varepsilon)x^*} \right]$$

$$N [\delta y^* + (1 - 2\varepsilon)x^*] = N_B \delta \left[\coth \beta \delta y^* + (1 - 2\varepsilon)x^* - \frac{1}{\beta \delta y^* + (1 - 2\varepsilon)x^*} \right] \quad (3.59)$$

$$+ N_A (1 - 2\varepsilon) \left[\coth \beta \delta x^* + (1 - 2\varepsilon)y^* - \frac{1}{\beta \delta x^* + (1 - 2\varepsilon)y^*} \right]$$

One can then come back to Equation 3.56, continue Laplace's method extracting the term $e^{Nf(x^*, y^*)}$ from the integral and taking the logarithm to obtain:

$$\log Z_b(\beta, \delta, \varepsilon) = Nf(x^*, y^*, \beta, \delta, \varepsilon) + \text{const.} + \mathcal{O}(N) \quad (3.60)$$

Now, with the partition function, we have the thermodynamical model and we can understand the behaviour of the system under different regimes of the parameters.

4

Results

We explore the consequences to the models presented in Section 3 and how the predictions we extract from them can be compared to empirical data or reasonable observations from contemporary societies.

4.1 Model 1

Although we could not find analytical solutions to Equation 3.16 in Section 3.1, they can be solved numerically by an iterative process as follows: let $u = (m, r, \zeta)$, then the set of equations is of the type $u = F(u)$, where F is given by Equation 3.16. An initial value u_0 is chosen and we iterate the map below until reaching a fixed point.

$$u_t = (1 - \alpha)u_{t-1} + \alpha F(u_{t-1}) \quad (4.1)$$

Since convergence was not a problem, no attempt at optimizing α was made. The result can be checked by choosing different starting points u_0 . The results are shown in Figure 4.1.

It can be seen that there is a phase transition depending on the parameters $\beta\nu$ and γ . We investigate further this transition by looking at the phase diagram in Figure 4.2.

Figure 4.2 shows that, for fixed β , the phase border can be crossed by increasing the value of γ . This seems paradoxical, since larger γ is associated to a larger norm of the covariance matrix (that is, a higher uncertainty in Equation 2.36). The explanation for this comes from

Figure 4.1: Solutions of Equation 3.16. Top: The normalization of the MF distribution ζ as a function of the social pressure and number of neighbors ($\beta\nu$). Bottom: Magnetization m . The other order parameter r has a similar behavior to m .

the fact that the gradient of the evidence (also called *modulation function* F_{mod}), which determines the dynamics in Equation 3.1, increases in magnitude with γ when both agents concur. This is shown in Figure 4.3

High γ agents rely not only on the novelty brought by disagreement but also learn from corroborating examples. For low γ , agents learn primordially from the novelty of disagreement. Therefore high γ agents will keep on learning even after there is agreement on an issue, resulting in a more ordered society. This same behavior was found in previous works (see e.g. Alves 2015; Caticha, Cesar, and Vicente 2015; Alves and Caticha 2016)

In Section 4.2 we show that another parameter, $\tilde{\beta}$, has a similar role in defining the “corroboration level” of a society, and we will give another interpretation closely related to this one.

We can also change variables in our mean field probability distribution. This is useful because the inner representation $\mathbf{J}_i$ is not readily accessible to the experimentalist, whereas the opinion field h in some applications might be:

$$P(h|\varepsilon, \gamma, \beta\nu) = \int d\mathbf{J} \delta\left(\frac{\mathbf{J} \cdot \mathbf{z}}{\sqrt{K}} - h\right) P_{MF}(\mathbf{J}) \quad (4.2)$$

$$= \frac{1}{C} (1 - h^2) \left[\varepsilon + (1 - 2\varepsilon) \Phi\left(\frac{r}{\gamma}h + \frac{m}{\gamma} \text{sign } h\right) \right]^{\beta\nu\gamma^2} \quad (4.3)$$

Figure 4.2: Phase diagram in the space $\gamma \times \beta\nu$. A phase transition separates an ordered from a disordered phase as signaled by the value of the order parameter m . Here, the value of ε was 0.1 and as it grows towards 0.5 the ordered phase decreases.

Figure 4.3: Modulation function/gradient of the evidence for different values of γ . The noise ε is fixed at 0.2, but changing to other values between 0. and 0.5 does not change qualitatively the figure

where m and r remain the expected values defined before: $m = \langle h \rangle$ and $r = \langle \text{sign } h \rangle$. Now we can also compute other order parameters, such as the variance $v_m = \langle h^2 \rangle - \langle h \rangle^2$.

The computational results we found are presented in Figure 4.4 and Figure 4.5.

Figure 4.4: Probability distribution $P(h|\varepsilon, \gamma, \beta\nu)$ for a set of values for the model's parameters.

Figure 4.5: Variance of the field h with respect to selected values of $\beta\nu$ and γ

We note that the distribution $P(h|\varepsilon, \gamma, \beta\nu)$ (Figure 4.4) has a discontinuity in $h = 0$ due to the sign h term, which is a source of complication when trying to fit empirical data. This is one of the main reasons we discuss a different model in Section 3.2, and 4.2

A last point to observe is that higher values of ε shift the phase boundary in Figure 4.2 upwards. This is because there is more noise in the interaction between agents therefore

it is more difficult to establish agreement. However, for values $\varepsilon \geq 0.5$, the noise is large enough that the system yields no stable solution. This is the first idea that led us to pursue a different model with explicit antiferromagnetic interactions in Section 3.3, and 4.3.

4.2 Model 2

Recall that in Section 3.2 we had two important results: Equation 3.19, which describes the model's main prediction $P(h|\tilde{\beta})$; and Equation 3.28 which enables us to estimate the parameter $\tilde{\beta}$ given an histogram of opinions $\{h_i\}$ for a given population pa and a fixed Zeitgeist $\mathcal{Z}$.

The Moral Foundations Questionnaire (see Graham, Haidt, and Nosek 2009) has answers given from subjects to 30 different questions in a scale from 0 to 5 assessing the importance the attribute to a certain statement aligned with one of the moral foundations. Because of this format, we had to perform a first processing on the data. We grouped the responses a subject gave to a moral foundation into a mean value of the answers aligned with that foundation. After that we normalized the 5-vector of an individual to 1. From that preprocessed data from the MFQ we could then obtain values for the moral vectors $\{\mathbf{J}_i\}_{i=1}^N$.

To have a dataset comparable to the model - which predicts only the distribution of opinions h of the agents -, we must make an additional hypothesis on how to choose the issue $\mathcal{Z}$ also from MFQ. Then the inner product between the chosen $\mathcal{Z}$ and $\{\mathbf{J}_i\}$ give us $\{h_i\}$. In (Vicente et al. 2014) they chose $\mathcal{Z}$ as the mean vector of Conservative respondents. Here we test 4 different hypothesis: mean of Conservatives $\mathcal{Z}_C$, mean of Liberals $\mathcal{Z}_L$, mean of Moderates $\mathcal{Z}_M$ and mean of all agents $\mathcal{Z}_T$. All the vectors were normalized $\|\mathcal{Z}\| = 1$.

We can see from Figure 4.6, as mentioned in (Haidt 2007), that one can differentiate through all 7 political affiliations (although here we show only Conservatives, Moderates and Liberals) by looking at the relevance each group attributes to each moral dimension. Liberals tend to give more importance to Harm and Fairness whereas Conservatives give almost equal importance to all 5 dimensions (slightly more to Ingroup, Authority and Purity dimensions). Additionally, the mean vector for the whole population tends somewhat more towards the Liberals due to a greater number of respondents with this political affiliation (pa).

The different estimates for $\tilde{\beta}$ for each of the Zeitgeists presented above were evaluated using Equation 3.28 and can be seen in Figure 4.7.

One can note that the choice of the Zeitgeist matters greatly when searching for a distribution $P(h|\tilde{\beta})$ that matches the data. More specifically, one can see from Figure 4.7 that the same group, for example $pa = 1$ (the Very Liberals), can have different values of $\tilde{\beta}$ for

Figure 4.6: The 4 different Zeitgeist hypothesis shown for comparison. All extracted from MFQ's respondents data.

Figure 4.7: Different posteriors for $\tilde{\beta}$ given the chosen Zeitgeist and the political affiliation of the responses $\{h\}$. Since the distributions are sharp, the error bars in the y -axis fall inside the markers (which are centered around the mean values)

different choices of the Zeitgeist.

The comparison of the fit $P(h|\tilde{\beta})$ with the histograms $\{h_i\}$ can be seen at Figure 4.8. This figure provides further insight: the model describes more accurately the data when the estimate values for $\tilde{\beta}$ are higher. This is probably because back in Equation 3.19 we assumed positive opinions $h > 0$, so we selected a model that better describes a consensus situation. The lower $\tilde{\beta}$ values, on the other hand, describe communities with a greater variance of opinions and less consensus.

Another information we can extract from the figures shown is that lower $\tilde{\beta}$ values appear when we choose Zeitgeists of Conservatives or Liberals, while the fit using Zeitgeists of Moderates or of the Total Population yield predictions more comparable with the data. We believe that - being the Conservatives and Liberals far from each other on the pa spectrum - choosing a Zeitgeist along one of those groups forces the other group to have widely spread opinions. It is a geometric feature of MFQ's data.

Figure 4.8: The histogram of opinions h for a given pa group considering an specific Zeitgeist and the corresponding best fit of the model $P(h|\tilde{\beta})$ given the data

From the analysis presented above, we propose 2 ideas for future models:

1. The choice of the question being discussed by the agents, the Zeitgeist, is a highly relevant characteristic of the model presented, and is not a feature we can extract

from the Moral Foundations data alone. One must have additional hypothesis, which ought to be validated by other means.

2. The model's parameter $\tilde{\beta}$, more than a characterization of the Liberal-Conservative spectrum, as was thought before, is a description of the extremism of a community. If one takes a Zeitgeist more aligned with the moral beliefs $\mathbf{J}_i$ of one political group than of other groups, then the distribution of opinions of the former will appear as more polarized than distributions from other political groups, and it will also have less room for a change of opinion. This behaviour, which is described by high value of $\tilde{\beta}$, constitutes what we call an "extremist" group.

4.3 Model 3

In Section 3.3 we presented a model for a bipartite society that presented a balance between in-group ferromagnetism and inter-group antiferromagnetism. This lead us to Equations 3.58, 3.59, a set of consistent relations we needed to solve to find the stable solutions (x^*, y^*) corresponding to the maxima of $f(x, y)$.

We repeat the results below for completeness:

$$f(x, y) = \frac{N_A}{N} \log \left(\frac{\sinh \beta \delta x + \beta(1 - 2\varepsilon)y}{\beta \delta x + \beta(1 - 2\varepsilon)y} \right) + \frac{N_B}{N} \log \left(\frac{\sinh \beta \delta y + \beta(1 - 2\varepsilon)x}{\beta \delta y + \beta(1 - 2\varepsilon)x} \right) \quad (4.4)$$

$$- \frac{\beta \delta}{2} (x^2 + y^2) - \beta(1 - 2\varepsilon)xy$$

Before solving the system it is useful to have a proper interpretation to x^* and y^* . To do that we can insert source terms $H_A \sum_{i \in A} h_i$ and $H_B \sum_{j \in B} h_j$ in the Hamiltonian (Equation 3.44):

$$\mathcal{H} = -\frac{\delta}{N} \sum_{\langle i, j \rangle \in A} h_i h_j - \frac{\delta}{N} \sum_{\langle i, j \rangle \in B} h_i h_j - \frac{(1 - 2\varepsilon)}{N} \sum_{\substack{i \in A \\ j \in B}} h_i h_j - H_A \sum_{i \in A} h_i - H_B \sum_{j \in B} h_j \quad (4.5)$$

This modifies the free energy to:

$$f(x, y) = \frac{N_A}{N} \log \left(\frac{\sinh \beta H_A + \beta \delta x + \beta(1 - 2\varepsilon)y}{\beta H_A + \beta \delta x + \beta(1 - 2\varepsilon)y} \right) + \frac{N_B}{N} \log \left(\frac{\sinh \beta H_B + \beta \delta y + \beta(1 - 2\varepsilon)x}{\beta H_B + \beta \delta y + \beta(1 - 2\varepsilon)x} \right)$$

$$- \frac{\beta \delta}{2} (x^2 + y^2) - \beta(1 - 2\varepsilon)xy \quad (4.6)$$

And give us the following partition function:

$$\begin{aligned} \log Z_B = N_A \log \left(\frac{\sinh \beta(\delta x^* + (1 - 2\varepsilon)y^* + H_A)}{\beta(\delta x^* + (1 - 2\varepsilon)y^* + H_A)} \right) \\ + N_B \log \left(\frac{\sinh \beta(\delta y^* + (1 - 2\varepsilon)x^* + H_B)}{\beta(\delta y^* + (1 - 2\varepsilon)x^* + H_B)} \right) - \frac{\beta\delta}{2} ((x^*)^2 + (y^*)^2) - \beta(1 - 2\varepsilon)x^*y^* \end{aligned} \quad (4.7)$$

Differentiating Equation 4.7 with respect to H_A we obtain:

$$\left. \frac{\partial \log Z}{\partial H_A} \right|_{H_A=0} = N_A \beta \left[\tanh(\beta\delta x^* + (1 - 2\varepsilon)y^* + \beta H_A) - \frac{1}{\beta\delta x^* + (1 - 2\varepsilon)y^* + H_A} \right] \quad (4.8)$$

On the other hand, differentiating directly the definition of the partition with respect to the source lead us into a mean value we associate with a magnetization:

$$\begin{aligned} \left. \frac{\partial \log Z}{\partial H_A} \right|_{H_A=0} &= \frac{1}{Z} \frac{\partial}{\partial H_A} \int dh \exp(-\beta\mathcal{H}) \\ &= \frac{1}{Z} \int dh \left(\beta \sum_{i \in A} h_i \right) \exp(-\beta\mathcal{H}) = \beta \left\langle \sum_{i \in A} h_i \right\rangle = \beta N m_A \end{aligned} \quad (4.9)$$

where we defined $m_A \equiv \frac{1}{N} \left\langle \sum_{i \in A} h_i \right\rangle$ and $m_B \equiv \frac{1}{N} \left\langle \sum_{j \in B} h_j \right\rangle$.

Now we can take Equations 3.58, 3.59:

$$\begin{aligned} N[\delta x^* + (1 - 2\varepsilon)y^*] &= N_A \delta \left[\coth \beta\delta x^* + (1 - 2\varepsilon)y^* - \frac{1}{\beta\delta x^* + (1 - 2\varepsilon)y^*} \right] \\ &\quad + N_B(1 - 2\varepsilon) \left[\coth \beta\delta y^* + (1 - 2\varepsilon)x^* - \frac{1}{\beta\delta y^* + (1 - 2\varepsilon)x^*} \right] \end{aligned} \quad (4.10)$$

$$\begin{aligned} N[\delta y^* + (1 - 2\varepsilon)x^*] &= N_B \delta \left[\coth \beta\delta y^* + (1 - 2\varepsilon)x^* - \frac{1}{\beta\delta y^* + (1 - 2\varepsilon)x^*} \right] \\ &\quad + N_A(1 - 2\varepsilon) \left[\coth \beta\delta x^* + (1 - 2\varepsilon)y^* - \frac{1}{\beta\delta x^* + (1 - 2\varepsilon)y^*} \right] \end{aligned} \quad (4.11)$$

and compare with Equation 4.9 and Equation 4.8 to find that the stable solutions x^* and y^* are the magnetizations for each community:

$$x^* = m_A \quad y^* = m_B \quad (4.12)$$

Now we perform once again the iterative process described in Equation 4.1 for several values of the parameters $(\beta, \delta, \varepsilon)$ and obtain the fixed points of the system. Once again optimizing α was unnecessary because the convergence was not a problem.

We considered only the situation in which the populations have the same size, that is, $N_A = N_B$.

Figure 4.9: The phase diagram when $\beta = 10.0$ and $N_A = N_B$. The initial points for the iterative algorithm were $x_0 = 0.9, y_0 = 0.1$

From Figure 4.9, beyond the expected behavior of an all-agreeing phase and a disagreeing one, we also note that there is a region where both groups present zero magnetization. This can be seen also from the free energy plot below (Figure 4.10), and is due to the interplay between agreement and disagreement in the groups' interactions: it is a region of the $(\beta, \delta, \varepsilon)$ -space inside which the agents cannot decide whether it is best to agree with their peer group or to disagree with the foe group, so the society enters a frustrated state. In statistical mechanics terms: both strategies (agreeing with my group or disagreeing with the other one) don't succeed in minimizing the overall energy of the system, so the only solution is to have zero magnetization throughout the society.

We also present the other 2 behaviours, expected from previous models: consensus in Figure 4.11 and polarization in Figure 4.12. Note that, depending on the initial choice of alignment of the communities (x_0, y_0) , the society (or, the iterative algorithm) converges to different and opposing states. This is not a statement about the dynamics of the model, but only on which group would align in favor of or against a given opinion (e.g. in one hypothetical situation community A could be in favor ($\sigma = +1$) of $\mathcal{Z}$ and community B against it ($\sigma = -1$), whereas under another initial condition the communities would change roles).

Figure 4.10: Free energy $f(x, y)$ landscape when $\beta = 10.0, \delta = 0.25, \varepsilon = 0.6$ and $N_A = N_B$. This is a region in which there is no consensus among the communities

We also show in Figure ?? the free energy plot at $\varepsilon = 0.5$. Considering high values for δ , since the society is at the transition between consensus and polarization, both scenarios can be observed depending on the choice of the initial points (x_0, y_0) .

Figure 4.11: Free energy $f(x, y)$ landscape when $\beta = 10.0, \delta = 0.8, \varepsilon = 0.2$ and $N_A = N_B$. This is a region in which there is overall consensus among the society

Figure 4.12: Free energy $f(x, y)$ landscape when $\beta = 10.0, \delta = 0.8, \varepsilon = 0.8$ and $N_A = N_B$. This is a region in which there is internal consensus on each community but they disagree with one another.

5

Conclusion and Perspectives

On this dissertation we were interested in studying some features of agents learning and interaction in society. We pursued this through the motif of morality, known to be a highly relevant element of discussion in society, and which provides good experimental and theoretical scaffold to develop our models.

To pursue that goal we presented 3 different models derived from the same maximum entropy inference procedure, each appealing to some level of description of the problem at hand.

5.1 Discussion of the results

In the first model we tried to be as faithful as possible to the results obtained from entropic dynamics (Section 2.5). This model takes an entropic inference model of what a learning procedure of a social agent should be and extrapolates it to a society of agents.

We completed a mean field approximation, but not without some compromises to the model. We found a phase transition between an ordered phase and a disordered one, representing two different states the society of agents could be.

The phases can be characterized by parameters γ and $\beta\nu$. Respectively, they are the variance of distribution $Q_n(\mathbf{B})$ - the entropic dynamics inference being made in Section 2.5 - and the product of the lagrange multiplier β of the statistical mechanical model in Section 3.1 with the effective number of peers ν of the mean field approximation.

Higher γ and higher $\beta\nu$ being the consensus society (ordered society, the one with non-zero magnetization). Lower values for both give us a society without any consensus (disordered, with zero magnetization).

Now, in light of the results obtained, we can better interpret the parameters:

- γ represents the uncertainty when dealing with incomplete information about the classifier vector $\mathbf{B}$. Higher γ is related to an agent being more open to learning from errors and adapting to the opinion of their peers. We saw that in Figure 4.3;
- $\beta\nu$ came from the statistical mechanical model and the mean field approximation: it arises from β , the pressure that constraints the learning to an specific behaviour, and ν , the effective number of peers each agent has. Higher $\beta\nu$ thus represents a greater peer pressure forcing the agent to adapt to the overall environment, therefore it is also a polarizing force.

Now we have a characterization of the type of agent that contributes to a society in which consensus reigns: first, it is an agent open to learning from corroboration of opinions as much as with novelty. This behaviour creates a kind of “echo chamber”, because the agents emit similar opinions repeatedly and change their internal beliefs (the moral vectors $\mathbf{J}$) even in situations where there is already agreement between the interacting pair; secondly, the typical agent in consensus societies is also one highly connected, because the strength of the pressure peers exert over each other grows with the number of connections made, ν .

We are convinced that the other kind of society described by this model, the one without magnetization, does not exist in reality (at least, not usually and not for a long period of time). This is because it should be unsustainable to have a society in which all the agents do not share common values about the only thing that is discussed between them. In that scenario the agents have strictly held opinions (low γ) and are isolated from one another, not being pressured much by their peers (low $\beta\nu$). In the third model we look more deeply into this problem from a perspective of a *normless* society.

The second model, built upon the work by (Vicente et al. 2014), was dedicated to establishing a comparison to data available from the MFQ. We offered a simple bayesian method to calculate the best parameter $\tilde{\beta}$ to a given dataset using Gamma distributions, and we challenged some of the assumptions presented before in the literature.

We showed how the choice of zeitgeist from the MFQ’s data is not an immediate nor trivial one, and that one must have additional hypothesis to support that choice. The choice of a zeitgeist aligned with conservatives may have overall good support from experimental evidence, but one can think in a hypothetical scenario in which the most discussed topic may be aligned with liberals, or with any other direction of the space. In that case we could not think of conservatives as behaving “conservatively”, since the group would have more

spread opinions and would accept novelty more than corroboration.

In sum, what we propose on this dissertation is that what has been called before a feature of “conservatism” must be re-framed to “*agents with strong agreement with the zeitgeist*”. For example, agents with a strong agreement to the zeitgeist are the ones who give the same importance to novelty or to corroborations, not conservatives (in the case when the zeitgeist is mostly aligned with conservatives then the conservatives behave like that).

We like to call those groups **extremists** because their behaviour is similar to what we would think of those in society. Imagine a polemic contemporary topic being discussed in society. We note the discussion and the groups in it being typically:

- highly polarized, is as if there are only two positions one can hold;
- both sides do not have much variability on their opinions, either you agree with them or you are from “the other group”;
- both sides focus much on posting on and propagating through their social networks situations and arguments that corroborate their opinions (much more than those that challenge their beliefs).

In the third model we tried to describe the framing presented above: two opposite groups discussing some issue and the struggle between agreeing with my trusted peers and disagreeing with my (possibly) untrusted peers.

We saw that the two obvious outcomes were present: the “*polarized society*” and the “*consensus society*”. However, the model also presented a third situation not entirely expected: a “*frustrated society*”. The phase diagram in Figure 4.9 shows the parameters’ regime for each of the phases. We describe them below

- High δ and high ε : *polarized society*

This is because we have high intra-group support for agreeing with my trusted peers but we also have strong inter-group antiferromagnetic interactions, which polarizes the society;

- High δ and low ε : *consensus society*

Now, with low modulation in the antiferromagnetic interactions, the agents minimize the energy by simply agreeing throughout the whole society;

- Low δ and low ε : *frustrated society*

The non-trivial case is one in which there is weak intra-group and inter-group interactions. This leads into an situation which only minimizes the energy by depolarizing the society

The concepts of societies in which there is overall consensus or polarized societies are easy

to grasp, but the concept of a frustrated one deserves more attention. Here we start to understand some of the results mentioned in Section 2.1 in view of the model presented:

- first, (Sherif 1937) noted that norms established in a group were important for a subject to retain a certain opinion in a next trial;
- secondly, (Deutsch and Gerard 1955) made a distinction between “normative” and “informational” types of influence, the first kind being the stronger of the two;
- and lastly, (Abrams et al. 1990) emphasized the importance of categorization: perceiving to which group I and my peers belong, and adapting accordingly.

All the studies mentioned highlight the importance of norms in a group to guarantee its cohesiveness and to protect the groups’ opinions from another group. Without strong norms inside a group it is easy for an external, rival, group to inject concepts and beliefs and co-opt the first group. It is also a strong set of norms that gives people the sense of belonging to something, and stimulates them to work towards some collective ideal.

In our third model, a frustrated society is a society in which the communities are not well established, there is no categorization, neither defining who is my friend nor who is my enemy. This is in contrast with the results found in the works mentioned, because in those works there were clear categories, and groups of individuals influencing other subjects (or not, when these did not trust the peers to be from their group).

We believe that this behaviour in our model can be attributed to a society with a weak set of norms. This can be seen in both variables: δ defines the strength of the moral norm required to be categorized as a member of one group; and ε the strength of the moral norm of *not* being part of the *other* group. If both variables have low values, then the agents perform poor categorization and we have zero magnetization throughout the society. This is a worse situation for the community than a polarized society, because now the group is vulnerable to a hypothetical third party to arrive and “attack” its weak moral or cultural identity.

5.2 Final Remarks

We presented several small results that advance the understanding of the agent-models that can be developed with Maximum Entropy and the framework shown here, but there are many more ideas worth pursuing using those models or even inventing others.

For example, we have reasons to believe that it is possible to find regimes of the parameters in the third model that yield an analytical solution to the expression for the partition function. One could also investigate the behaviour of the system when $N_A \neq N_B$, and see whether the model agree with the effects of a majority group encountered experimentally

by (Asch 1956). Those two ideas were not pursued due to the time constraint the student was subject to.

As for new initiatives, we name some perspectives for the future:

- There has been some research in our group (possibly Alves and Caticha 2018) to understand situations in which the society discusses more than one question. This could lead into richer structures of organization in the society;
- Another possibility that was investigated by the student but so far found inconclusive results is the comparison of the peer pressure parameter in the first model, β , with the *tightness* measure presented in (Gelfand et al. 2011). There has been some work on this by (Cesar 2014), but we think that our new concept of extremists may help develop further insights on this topic;
- Closely related to the last point, one could also do the inference performed in the second model taking into account the other variables the MFQ had and we ignored on a first approach, such as country, age, and the variability inside the 30 questions. Now, with a model for $P(g|\text{country}, pa, \text{question})$, one could try to generalize something about the differences between countries, the political groups and the questions being asked. This would require a different preprocessing of the MFQ data, instead of taking the mean value of several questions at once;
- An idea that emerges from the topic of norms and categorization discussed in the third model is the existence of *convincing strategies* that can co-opt one group into thinking like another group. We believe this is an important problem worth investigating due to several relevant situations that can be cast into this frame, namely: external agent influencing an election, mass opinion manipulation, cultural, moral and religious colonialism, among others;
- One last idea we would like to propose is a re-framing of the problem: instead of performing inference over the learning of one agent, and extrapolating it to a society of agents, one could try to think on an inference over the society as a whole. We believe this could provide alternative intuitions on the evolution of the system.

We believe this line of work is only flourishing, with many possible avenues still worth exploring. The problem is complex, rich in details, and difficult, which only stimulates further research in the area. The tools presented have proved to have great descriptive power and reasonable interpretations to deal with the problem at hand.

References

- Abrams, Dominic, Margaret Wetherell, Sandra Cochrane, Michael A. Hogg, and John C. Turner. 1990. "Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization." *British Journal of Social Psychology* 29 (2): 97–119. <https://doi.org/10.1111/j.2044-8309.1990.tb00892.x>.
- Alves, Felipe. 2015. "Quebra de Simetria Espontânea, Limites Cognitivos e Complexidade de Sociedades." MSc Thesis, São Paulo: Universidade de São Paulo. <https://doi.org/10.11606/D.43.2015.tde-27042015-101234>.
- Alves, Felipe, and Nestor Caticha. 2016. "Sympatric multiculturalism in opinion models." In *AIP Conference Proceedings*, 1757:060005. <https://doi.org/10.1063/1.4959064>.
- . 2018. "Entropic Dynamics of Distrust and Opinions of Interacting Agents." *In Preparation*.
- Asch, Solomon E. 1956. "Studies of independence and conformity: I. A minority of one against a unanimous majority." *Psychological Monographs: General and Applied* 70 (9): 1–70. <https://doi.org/10.1037/h0093718>.
- Axelrod, Robert. 1997. "The Dissemination of Culture." *The Journal of Conflict Resolution* 41 (2): 203–26. <https://doi.org/10.1177/0022002797041002001>.
- Axelrod, Robert, and W. Hamilton. 1981. "The evolution of cooperation." *Science* 211 (4489): 1390–6. <https://doi.org/10.1126/science.7466396>.
- Baumeister, Roy F., and Mark R. Leary. 1995. "The need to belong: desire for interpersonal attachments as a fundamental human motivation." *Psychological Bulletin* 117 (3): 497–529. <https://doi.org/10.1037/0033-2909.117.3.497>.
- Braver, Todd S, D M Barch, J R Gray, D L Molfese, and A Snyder. 2001. "Anterior cingulate cortex and response conflict: Effects of frequency, inhibition and errors." *Cerebral Cortex* 11 (9): 825–36. <https://doi.org/10.1093/cercor/11.9.825>.
- Brown, Jennifer L., David Sheffield, Mark R. Leary, and Michael E. Robinson. 2003. "Social support and experimental pain." *Psychosomatic Medicine* 65 (2): 276–83. <https://doi.org/10.1097/01.PSY.0000030388.62434.46>.
- Buck, R W, and R D Parke. 1972. "Behavioral and physiological response to the presence of a friend or neutral person in two types of stressful situations." *Journal of Personality and Social Psychology* 24: 143–53.
- Castellano, Claudio, Santo Fortunato, and Vittorio Loreto. 2009. "Statistical physics of social dynamics." *Reviews of Modern Physics* 81 (2): 591–646. <https://doi.org/10.1103/RevModPhys.81.591>.
- Caticha, Ariel. 2012. *Entropic Inference and the Foundations of Physics*. <http://www.albany.edu/physics/ACaticha-EIFP-book.pdf>.

———. 2017. “Entropic Dynamics: Quantum Mechanics from Entropy and Information Geometry.”

Caticha, Nestor, Jonatas Cesar, and Renato Vicente. 2015. “For whom will the Bayesian agents vote?” *Frontiers in Physics* 3 (25): 1–14. <https://doi.org/10.3389/fphy.2015.00025>.

Caticha, Nestor, and Renato Vicente. 2011. “Agent-based Social Psychology: From Neurocognitive processes to Social data.” *Advances in Complex Systems* 14 (05): 711–31. <https://doi.org/10.1142/S0219525911003190>.

Cesar, Jonatas. 2014. “Mecânica estatística de sistemas de agentes bayesianos: aplicação à teoria dos fundamentos morais.” PhD Thesis, São Paulo: Universidade de São Paulo. <https://doi.org/10.11606/T.43.2014.tde-30102014-090629>.

Cox, Richard T. 2001. *Algebra of Probable Inference*.

Dehaene, Stanislas. 2018. “The Error-Related Negativity, Self-Monitoring, and Consciousness.” *Perspectives on Psychological Science* 13 (2): 161–65. <https://doi.org/10.1177/1745691618754502>.

Deutsch, Morton, and Harold B. Gerard. 1955. “A study of normative and informational social influences upon individual judgment.” *The Journal of Abnormal and Social Psychology* 51 (3): 629–36. <https://doi.org/10.1037/h0046408>.

Eisenberger, Naomi I. 2012. “The pain of social disconnection: Examining the shared neural underpinnings of physical and social pain.” *Nature Reviews Neuroscience* 13 (6): 421–34. <https://doi.org/10.1038/nrn3231>.

Eisenberger, Naomi I., and Matthew D. Lieberman. 2004. “Why rejection hurts: A common neural alarm system for physical and social pain.” *Trends in Cognitive Sciences* 8 (7): 294–300. <https://doi.org/10.1016/j.tics.2004.05.010>.

Eisenberger, Naomi I., Matthew D. Lieberman, and Kipling Williams. 2003. “Does Rejection Hurt? An fMRI Study of Social Exclusion.” *Science* 302 (5643): 290–92. <https://doi.org/10.1126/science.1089134>.

Gehring, William J., Brian Goss, Michael G.H. Coles, David E. Meyer, and Emanuel Donchin. 1993. “A Neural System for Error Detection and Compensation.” *Psychological Science* 4 (6): 385–90. <https://doi.org/10.1111/j.1467-9280.1993.tb00586.x>.

Gehring, William J., Brian Goss, Michael G. H. Coles, David E. Meyer, and Emanuel Donchin. 2018. “The Error-Related Negativity.” *Perspectives on Psychological Science* 13 (2): 200–204. <https://doi.org/10.1177/1745691617715310>.

Gelfand, Michele J., J. L. Raver, L. Nishii, L. M. Leslie, J. Lun, B. C. Lim, L. Duan, et al. 2011. “Differences Between Tight and Loose Cultures: A 33-Nation Study.” *Science* 332 (6033): 1100–1104. <https://doi.org/10.1126/science.1197754>.

Gilligan, Carol, Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. 2012. “Mapping the Moral Domain.” *Journal of Personality and Social Psychology* 101 (2): 366–85. <https://doi.org/10.1037/a0021847.Mapping>.

Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. “Moral Foundations Theory.” In *Advances in Experimental Social Psychology*, 47:55–130. <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>.

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. “Liberals and conservatives rely on different sets of moral foundations.” *Journal of Personality and Social Psychology* 96 (5): 1029–46. <https://doi.org/10.1037/a0015141>.

- Greene, Joshua D. 2009. "The cognitive neuroscience of moral judgment." *The Cognitive Neurosciences IV*, 1013–24. <http://xa.yimg.com/kq/groups/14661755/505292463/name/gazzaniga-greene+chap.68.pdf>.
- Haidt, Jonathan. 2001. "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological Review* 108 (4): 814–34. <https://doi.org/10.1037/0033-295X.108.4.814>.
- . 2007. "The new synthesis in moral psychology." *Science (New York, N.Y.)* 316 (5827): 998–1002. <https://doi.org/10.1126/science.1137651>.
- Haidt, Jonathan, and Craig Joseph. 2004. "Intuitive ethics: how innately prepared intuitions generate culturally variable virtues." *Daedalus* 133 (4): 55–66. <https://doi.org/10.1162/0011526042365555>.
- Haidt, Jonathan, and Selin Kesebir. 2010. "Morality." *Handbook of Social Psychology*, 797–832. <https://doi.org/10.1002/9780470561119.socpsy002022>.
- Holroyd, Clay B., and Michael G. H. Coles. 2002. "The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity." *Psychological Review* 109 (4): 679–709. <https://doi.org/10.1037//0033-295X.109.4.679>.
- Iyer, Ravi, Spassena Koleva, Jesse Graham, Peter H. Ditto, and Jonathan Haidt. 2012. "Understanding libertarian morality: the psychological dispositions of self-identified libertarians." *PloS One* 7 (8). Public Library of Science: e42366. <https://doi.org/10.1371/journal.pone.0042366>.
- Jaynes, Edwin T. 1957a. "Information Theory and Statistical Mechanics." *Physical Review* 106 (4): 620–30. <https://doi.org/10.1103/PhysRev.108.171>.
- . 1957b. "Information Theory and Statistical Mechanics. II." *Phys. Rev. E* 108 (2): 171–90.
- . 1965. "Gibbs vs Boltzmann Entropies." *American Journal of Physics* 33 (5). American Association of Physics Teachers: 391–98. <https://doi.org/10.1119/1.1971557>.
- . 1980. "The Minimum Entropy Production Principle." *Annual Review of Physical Chemistry* 31 (1): 579–601. <https://doi.org/10.1146/annurev.pc.31.100180.003051>.
- Jaynes, Edwin T., and G. Larry Bretthorst. 2003. *Probability theory: the logic of science*.
- Newton, Isaac. 1675. *Isaac Newton letter to Robert Hooke, 1675 [electronic resource]*. Edited by Isaac Newton and Robert Hooke. <https://digitallibrary.hsp.org/index.php/Detail/objects/9792>.
- Pizarro, David A., and Paul Bloom. 2003. "The intelligence of the moral intuitions: A comment on Haidt (2001)." *Psychological Review* 110 (1): 193–96. <https://doi.org/10.1037/0033-295X.110.1.193>.
- Schelling, Thomas C. 1969. "Models of segregation." *The American Economic Review* 59 (2): 488–93.
- . 1971. "Dynamic models of segregation." *The Journal of Mathematical Sociology* 1 (2): 143–86. <https://doi.org/10.1080/0022250X.1971.9989794>.
- Schrödinger, Erwin. 1944. *What is Life*.
- Sherif, Muzafer. 1937. "An Experimental Approach to the Study of Attitudes." *Sociometry* 1 (1/2): 90. <https://doi.org/10.2307/2785261>.
- Shore, J E, and R. W Johnson. 1980. "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross Entropy." *IEEE Transactions on Information Theory* IT-26 (1): 26–37. <https://doi.org/10.1109/TIT.1980.1056144>.

Simões, Lucas Silva, and Nestor Caticha. 2018. “Mean Field Studies of a Society of Interacting Agents.” In, edited by Adriano Polpo, Julio Stern, Francisco Louzada, Rafael Izbicki, and Hellinton Takada, 239:131–40. Springer Proceedings in Mathematics & Statistics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91143-4_13.

Sivia, Devinderjit, and John Skilling. 1998. *Data Analysis: A Bayesian Tutorial*. Vol. 40. 2.

Skilling, John. 1988. “The Axioms of Maximum Entropy.” In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, 173–87. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-3049-0_8.

Vicente, Renato, Alex Susemihl, João Pedro Jericó, and Nestor Caticha. 2014. “Moral foundations in an interacting neural networks society: A statistical mechanics analysis.” *Physica A: Statistical Mechanics and Its Applications* 400 (c): 124–38. <https://doi.org/10.1016/j.physa.2014.01.013>.

Yeung, Nick, Matthew M. Botvinick, and Jonathan D. Cohen. 2004. “The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity.” *Psychological Review* 111 (4): 931–59. <https://doi.org/10.1037/0033-295X.111.4.939>.