# COMPUTATIONAL MODELLING OF SOCIAL COGNITION AND BEHAVIOUR

By

Nikos C. Theodoropoulos

*A Dissertation Submitted in Partial Fulfillment of the*

*Requirements for the Degree of*

DOCTOR OF PHILOSOPHY

in

Experimental Psychology and Cognitive Neuroscience

Supervised by:

Adam Moore, Dr
University of Edinburgh

Nadia Gamboz, Associate Professor
Università degli Studi Suor Orsola Benincasa

Chris Lucas, Dr
The University of Edinburgh

The University of Edinburgh
Università degli Studi Suor Orsola Benincasa

# Acknowledgements

I cannot express enough thanks to my primary supervisor, Dr. Adam Moore, for his continued support and encouragement. Thank you for always being reassuring when the insecurities got stronger.

The completion of this project could not have been accomplished without the support of my secondary supervisor, Professor Nadia Gamboz. Thank you for giving me the necessary suggestions to better this study and for the warm and unforgettable welcome in Napoli!

I also want to thank my secondary supervisor, Dr Chris Lucas. Your technical knowledge made a difference in my project.

Above all, I want to give my deepest gratitude to my mother, and the rest of my family. Your encouragement when the times got rough are much appreciated and duly noted.

# Contents

# Chapter 1

# Introduction

## 1.1 Introduction

Philosophers have always been interested in asking moral questions, but social scientists have generally been more occupied with asking questions about morality. How do people differ with regards to their morality? How frequently are moral values inconsistent, thus resulting in internal conflicts? How likely are people to revise their moral beliefs? The aim of these questions is to explore moral reasoning and identify patterns of moral behaviour between people. Simultaneously, social scientists have moved beyond the exploration of small-scale, static snapshot of networks onto nuanced, data-driven analyses of the structure, content, and dynamics of large-scale social processes. This drives researchers to use far more elaborate tools, such as automated text analysis, online field experiments, mass collaboration, machine learning, and more generally computational modelling, to formulate and test theories (e.g., Evans & Aceves, 2016; Molina & Garip, 2019; Nelson, 2020; Salganik, 2019). It is fair to argue that social sciences are on the verge of a new era, an era in which computational methods and large-scale data are the primary tools/sources of gaining information and knowledge.

In this dissertation, I will focus on developing formal models of the cognitive dissonance involved in moral values conflicts within individuals, and how this might be reduced. I will also attempt to extend this to connect with research linking moral and political psychology. Then I will try to explain echo chamber development, as a socio-cognitive phenomenon, arising from dynamics described in chapters 2 and 3. Finally, I will focus on moral belief updating, as an alternate (class of) response(s) in chapter 6.

I try to explain these phenomena by bringing together cognitive and social theories. The three principal theories we build upon are Festinger's Cognitive Dissonance, Bandura's Moral Disengagement and Haidt's Moral Fountations Theory. As it is detailed in the forthcoming paragraphs, the union of these theories, alongside with computational modelling, sparks off some interesting hypotheses. We now go ahead and discuss why computational modelling is a powerful tool in social sciences, and then present a historical background for each of the aforementioned theories.

### 1.1.1 Computational social science

The term *computational social science* was coined in the last decades of the 20nth century within social science fields as well as science, technology, engineering, and mathematics (STEM) fields. In social sciences, the term initially referred to agent based modelling –or more generally, the usage of computer algorithms to simulate group behaviour using artificial agents (Macy & Willer, 2002; Bruch & Atwell, 2015). Relevant efforts yielded essential theoretical progress in social psychology research, network analyses, and various other topics (e.g., Baldassarri & Bearman, 2007; Centola & Macy, 2007; Watts, 1999). In contrary, in STEM disciplines almost any work that describes human behaviour is often marked as computational social science (e.g., Helbing et al., 2000; Pentland, 2015). Even though the majority of these works exploited exquisite ideas from physics and mathematics to analyse shared dynamics such as group behaviour, they were mostly detached from social science research (see McFarland et al., 2016) –in spite of attempts to synchronise them (e.g., Carley, 1991; Macy & Willer, 2002).

Recognising its various use-cases between diverse disciplines, Edelmann et al. (2020) provided the subsequent definition of the discipline: "Computational social science is an interdisciplinary field that advances theories of human behaviour by applying computational techniques to large datasets from social media sites, the Internet, or other digitised archives such as administrative records" (p. 62). This description emphasises sociology as novel data sources and methods should not be the only parts of computational social sciences, rather a crucial factor is its potential to yield new hypotheses of human behaviour or advance current theories of the social world. That is exactly how we are using computational methods in this dissertation: we exploit the power of this approach in terms of developing formalisable theoretical models that can be probed for insight (see, for example, chapter 2), particularly with respect to otherwise difficult to study *real word* phenomena, such as political polarisation and echo chambers formation (see also chapter 5). In the same vein, we use Edelmann et al.'s (2020) definition because we are more interested in (and the projects included in this dissertation evolve around) explaining human behaviour to advance social science theory, rather than predicting human behaviour for practical purposes (as in Macy, 2015; Watts, 2017). In other words, our research emphasis is firmly on *why* and *how*, with *what* and *when* playing a secondary role.

### 1.1.2 Cognitive Dissonance

Cognitive dissonance theory (Festinger, 1957; Festinger & Carlsmith, 1959; Festinger, 1962) is one of the most prominent theories in social psychology (Jones, 1985). It has stimulated numerous insights across many research topics, such as drivers of behaviour and ideology, the internalization of beliefs, the consequences of choices, the impacts of disagreements between people, and more (Harmon-Jones & Mills, 2019).

Dissonance theory initially proposed that any pair of cognition –behaviour, attitudes, beliefs– can be either consonant, dissonant, or unrelated to each other. Two or more cognitions are *consonant* if one follows from the other, *dissonant* if the obverse (i.e., opposite) of one follows from the other, and irrelevant when the two cognitions are unconnected. For example, "one should protect innocent people" and "protecting innocent people makes you a good person" are two consistent cognitions, while "all lives matter" and "enemies must be killed" are two inconsistent cognitions, and last, "one should not abuse animals" and "one should take the stairs when possible" are two irrelevant cognitions. When two cognitions (or an action and a cognition) are dissonant the person experiences a psychological discomfort that motivates action to achieve or restore consonance. The larger the strength of the dissonance[1], the larger is the force to decrease it.

Dissonance may be decreased by disengaging from one of the dissonant cognitions, adding new consonant cognitions or reducing the significance of dissonant cognitions. The possibility that a specific cognitive element will alter to decrease the discomfort is given by how resistant to change the element is (i.e., its epistemic centrality cf. Quine, 1951; or epistemic entrenchment, cf. Gärdenfors, 1988; see also Dubois and Prade, 1991). The smaller the resistance, the easier the cognition will alter upon request to decrease dissonance. The resistance to alter of a behavioural cognition relies upon the degree of pain or damage which have to be tolerated and the pleasure received from the behaviour. A frequently used example (also used by Festinger, 1957) may help building up some intuition about the theory. If an individual is keep smoking even after they are informed of the health threats of smoking, it is expected to feel discomfort caused by the conflict between their knowledge (i.e., smoking is bad) and their behaviour (i.e., smoking). Decreasing or quit smoking altogether is one way of decreasing dissonance (i.e., disengaging from their behaviour). Otherwise, they may decrease the discomfort by altering their knowledge regarding the impact of smoking on health and consider smoking as not having any damaging consequence on health (ie., reducing the importance of the dissonant cognition). Last, the smoker could search for affirmative properties of that behaviour and suppose that smoking decreases stress and prevents them from getting weight (i.e., increasing consonant cognitions).

Since it was first introduced over half a century ago, cognitive dissonance theory has kept giving rise to new studies, revisions, and controversy. The theory was posed in greatly abstract concepts, making it applicable to a wide range of psychological subjects ranging from behaviours, and attitudes, to beliefs, perceptions, and affects. Moreover, cognitive elements can be about oneself, some other individual or group, or about things in the environment. Instead of being relevant to one topic, the theory is relevant to

[1] When Festinger were using the term *dissonance* was referring both to the difference among cognitions and to psychological distress. These two notions are theoretically different and the former is now referred to as cognitive inconsistency or cognitive discrepancy, while the latter as dissonance or dissonance discomfort.

multiple topics. Its abstract terms also make it easier to establish links between other theories, such as moral disengagement theory (Bandura, 1999), which comes into play when an individual seeks methods to decrease the unpleasant state of cognitive dissonance they feel after having contradictory moral cognitions.

### 1.1.3 Moral Disengagement

Moral disengagement describes eight interconnected cognitive mechanisms that enable one to cognitively separate the moral component from an otherwise unprincipled act (Bandura, 1986, 1999, 1990b; Bandura et al., 1991), in order to rationalise engaging in it and deactivating the mechanism of self-condemnation (Fiske, 2018), which role -under a healthy and well functioning moral psychology- is to prevent people from committing unethical behaviours. Therefore, moral disengagement assumes an operation of cognitive reconstruing or reframing of harmful behaviour as being ethically acceptable without altering the behaviour or the moral standards (Bandura, 1999). The following paragraph provides a brief introduction to these strategies (the interested reader can refer to Bandura, 1999, for a more detailed description).

For instance, suppose Kya has moral standards that condemn theft. Suppose also that Kya took a book from a bookstore without buying it. Moral disengagement mechanisms could aid Kya interpret taking the book as not important (*distortion of consequences*), or think that everybody takes small stuff such as a book every now and then (*diffusion of responsibility*), or that taking the book is insignificant compared to other people who rob a bank (*advantageous comparison*), or that extortionate pricing by booksellers has forced her to theft in order to further her education (*displacement of responsibility*). She might also reason that in a broader perspective, being a well-educated individual is more significant than paying for a book (*moral justification*). She might even plan on returning the book after she finishes reading it, so really she was just ”borrowing” it (*euphemistic labelling*). She might reason that this bookstore is simply a big ruthless corporation where a missing book will go unnoticed (*dehumanisation*), or even deserves having the book stolen from it since it charges this much (*attribution of blame*). As this example demonstrates, there can be overlap between these strategies, with multiple strategies being engaged simultaneously to produce what might be called hybrid disengagement cognitions (e.g., note similarity between displacement of responsibility and attribution of blame).

Moral disengagement theory has laid the foundations for empirical studies over numerous disciplines and fields, such as children and adolescents development (e.g., Pornari & Wood, 2010; Gini et al., 2011; Obermann, 2013; Gini et al., 2014; Caravita et al., 2014), organisational behaviour (e.g., Claybourn, 2011; Duffy et al., 2012; C. Moore et al., 2012; Martin et al., 2014; T. R. Cohen et al., 2014; Christian & Ellis, 2014; Samnani et al., 2014; J. F. Johnson & Buckley, 2015), criminology (e.g., DeLisi et al., 2014; Cardwell et al., 2015), military psychology (e.g., Beu & Buckley, 2004; McAlister et al., 2006; Aquino et al., 2007), and sports psychology (e.g., Hodge & Lonsdale, 2011; Boardley & Kavussanu, 2011; Hodge et al., 2013). Moral disengagement tendency is linked to several negative behaviours, including criminality (e.g., Cardwell et al., 2015), aggression and bullying (e.g., Pornari & Wood, 2010; Gini et al., 2011; Obermann, 2013; Gini et al., 2014), workplace misconduct (e.g., Barsky, 2011; Duffy et al., 2012; C. Moore et al., 2012), and unethical behaviour generally (e.g., Detert et al., 2008; C. Moore et al., 2012), as well as several negative psychological states, including an increased ability to dehumanise others (e.g., Castano, 2008; Leidner et al., 2010; Waytz & Epley, 2012) and higher probability of endorsing violence toward them (e.g., Osofsky et al., 2005; McAlister et al., 2006).

There is a subtle link between moral disengagement theory and the cognitive dissonance reduction strategies. Cognitive dissonance occurs when the individual feels tension from holding two or more contradicting cognitions - in this case, between moral and immoral/unethical behaviours/decisions (R. A. Baron et al., 2015). This unpleasant state motivates attempts to reduce the dissonance by either updating/revising one of the dissonant cognitions, adding consonant cognitions, or decreasing the importance of one of the dissonant cognitions. The moral disengagement mechanisms allow for sidestepping internalised moral standards and engaging in immoral behaviour without attendant feelings of distress by adding more consonant cognitions, or decreasing the importance of one of the dissonant cognitions, but never dropping a dissonant cognition altogether. In other words, these mechanisms allow individuals to engage in immoral behaviour without having to revise or update their moral beliefs or moral self-image. Thus, this thesis treats moral disengagement strategies as an extension of two dissonance reduction tactics (i.e., adding consonant cognitions, decreasing the importance of one of the dissonant cognitions) introduced by cognitive dissonance theory.

Discussion of dissonant cognitions in the moral domain, however, necessitates the explication of moral beliefs/values. As it turns out, there is significant theoretical and empirical evidence for specific and interesting structure in human moral values and the beliefs based on them (Haidt & Joseph, 2004, 2007; Haidt & Graham, 2007; Haidt, 2012).

### 1.1.4 Moral foundations theory

The research of moral psychology is -broadly speaking- divided into two somewhat contradictory views. The first and more traditional view supports that moral judgement is the result of conscious, effortful reasoning and the emotions follow (Kohlberg, 1969; Piaget, 1965; Turiel, 1983; Killen, Smetana et al., 2005). The more recent years, this view became less popular leaving more space to the rival view which claims that moral thinking is something we do only after a spontaneous processes (passion/emotions/intuitions) have already directed us towards a judgment or decision, and we do so to get ready for social intercourses where we may be asked to defend our judgements to other individuals (i.e., persuade others, or defend oneself; Haidt, 2001; Hauser, 2006; Mikhail, 2000; Shweder and Haidt, 1993; see also Damasio, 1994).

One of the advocates of the later view is Moral foundations theory (MFT) (Haidt, 2001; Haidt & Joseph, 2004, 2007; Haidt et al., 2009; Graham et al., 2009) which tries to describe the roots of and variance in human moral reasoning based on innate, modular foundations, and supports that moral reasoning merely functions as a post-hoc rationalization of already shaped judgements (Haidt, 2001). In particular MFT proposes that individuals come equipped with a few *intuitive ethics*, or modules, that are formed from the evolutionary process as a response to adaptive challenges (Haidt & Joseph, 2004). It goes on to propose that each module is organised in advance of experience but cultural learning process shapes each of these modules. Moral values deviate because differing cultures utilise the "building blocks" presented by the foundations dissimilarly (Haidt & Joseph, 2004).

### The five moral foundations

Initially, (Haidt & Joseph, 2007) five candidates that can explain the variance in individual's moral values were identified: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation [2].

1. Care: protecting and caring for others, being kind; obverse of *harm.*
2. Fairness: being just and right based on shared rules: obverse of *cheating.*
3. Loyalty: protecting and caring for your own group/family/nation; obverse of *betrayal.*
4. Authority: endorsing to leadership and followership: obverse of *subversion.*
5. Sanctity: aversion of *disgusting* things and actions: obverse of *degradation.*

These five foundations can be clustered into two higher-level categories: 1) the individual-focused *individualising* category which consist of *care* and *fairness* foundations, and 2) the group-focused *binding* category which consists of *loyalty*, *authority*, and *sanctity* (Graham et al., 2009; Haidt & Graham, 2009). The empirical data supporting this categorising arises from patterns of relationships among the moral foundations observed with the Moral Foundations Questionnaire (Graham et al., 2009; Graham et al., 2011).

One further candidate for the list of the foundations exists: Liberty/oppression (Iyer et al., 2012). However, this foundation is not as well-researched as the initial five, and so, in the current work I will focus on the established five moral foundations, though

[2] For reasons of brevity, we will refer to the five foundations using only the left side of each foundation, e.g., care/harm will be referred to as care.

in principle, there is no reason that any of the current work cannot be extended to accommodate six foundations.

Moral foundations theory has given rise to several new studies, ideas, and controversy. Numerous researchers have given critiques of (Suhler and Churchland, 2011; Schein and Gray, 2018, for the answers to these critiques see Haidt and Joseph, 2011; Koleva and Haidt, 2012), and alternatives to (e.g., Rai & Fiske, 2011; Gray, Young et al., 2012; Verhulst et al., 2012; Janoff-Bulman & Carnes, 2013; Schein & Gray, 2018; Curry et al., 2019) the theory. Besides controversy, however, the theory has made valuable advancement in many fields with perhaps the most apparent being politics (Graham et al., 2009).

### Moral foundations theory applied to political cultures

Individuals' sensitivity to the moral foundations is related to their political ideology. Graham et al. (2009)[3] report four works exploring the moralities of political agents on the liberal/left - conservative/right continuum. In the first study, a wide worldwide sample of participants evaluated the moral importance of foundations-specific matters. In the second research, they explored lefts' and rights' moral judgements in response to either explicit or implicit political identities. In the third research, they derived more robust intuitive reactions by posing subjects ethical trade-off questions asking them how much they want to get paid to engage in foundations-violating behaviours. In the last research, they analysed sermon passages preached in leftist and rightist churches to check whether preachers in the diverse moral groups intuitively employed foundations-related language in distinct ways. The results were clear: liberals showed evidence of a moral values according mainly to the individualising foundations (care and fairness), while conservatives displayed a more uniform distribution of values, ideas, and concerns, involving all five foundations. See figure 1.1 for an overview of their results.

Moral foundations theory and its questionnaire have had an enormous impact on research on politics, but as every good theory, it has shortcomings. The unidimensional spectrum used in describing political ideology might fail capturing patterns of morality in

[3] Visit moralfoundations.org and www.YourMorals.org for more details on the questionnaire.

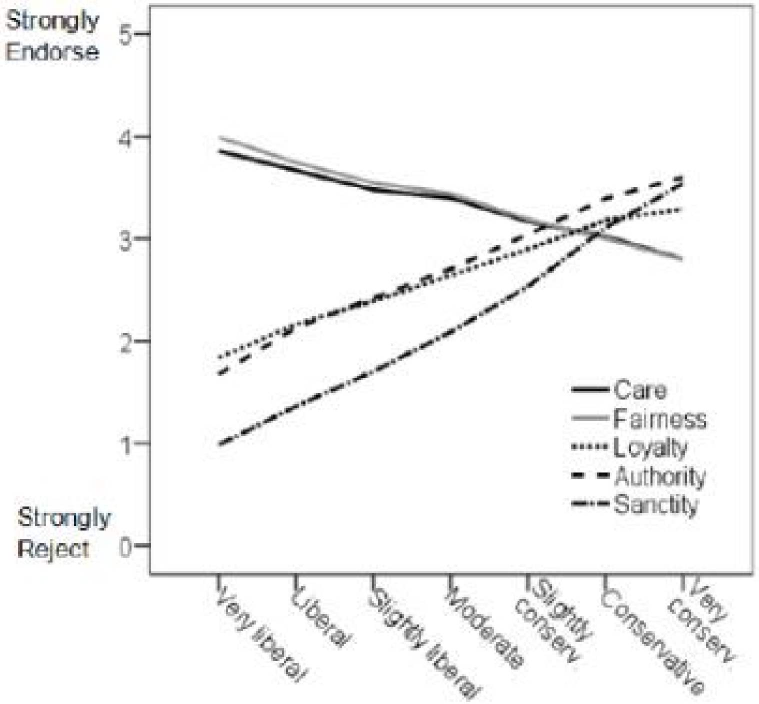


*Figure 1.1.* Importance of moral foundations over political identities, from study 1. Adapted from *Leftists and rightists rely on differing sets of moral foundations* by Graham, J., Haidt, J., and Nosek, B. A., 2009

more complex ideologies. In particular, relevant studies showed that variants of political orientations (i.e., variants of liberals and conservatives) value the foundations in different ways (Iyer et al., 2012; Weber & Federico, 2013). Still, this one dimensional political continuum (liberals-conservatives/left-right) is a handy estimation tool that has prognostic validity for voting behaviour and opinion on a large scope of matters (Jost, 2006), and as such, this is what I use in the current work as well.

Moral foundations theory is the link needed to bridge the cognitive dissonance and the moral disengagement theories to our research questions. If different political ideologies uphold different moral beliefs then there might be a difference in the frequency of these beliefs to be conflicting between the various political ideologies, which in turn drives to a potentially diverse rate of cognitive dissonance and then disengagement of the dissonant beliefs. In other words, a given political ideology might experience less or more dissonance as a function of their moral complexity. To explore the differences in the frequency of conflict among different political orientations, we use a theoretical framework using the help of graphical modelling (Lauritzen, 1996; Koller & Friedman, 2009).

### 1.1.5 Probabilistic graphical models

Probabilistic graphical models (PGMs) offer a succinct portrayal of a joint probability distribution of potentially numerous dimensions of variables. To do so it is utilising conditional independencies in the network of these variables; a network like this, with local in/dependencies is named a *graphical model*. PGM modelling is ingrained in probabilistic reasoning, questioning, as well as sampling (Koller & Friedman, 2009). A detailed explanation on PGM can be found from appendices A.1 to A.2.

Bayesian networks (BNs) are a particular subclass of PGM that represent targeted dependencies among variables capturing cause-to-effect relations. This feature renders BN an convenient modelling approach for social systems that involve and convey such relationships. Specifically, BNs have been used for social network analysis and the prediction of new connections and prespecified unit traits (e.g., teamwork potentials) (Koller & Friedman, 2009). They emphasise the leverage of BNs to explain uncertainty, noise, and incompleteness in the system. For instance, topological measures like *degree centrality* that is frequently used as an *importance* indicator, is susceptible to summarisations across partial or even erroneous data. In contrast, BNs are more flexiblle in that they enable estimates like importance to be predicted in a more data-dependent way. Koelle et al. (2006) offer a paradigm of merging topology-based system estimates with co-variate information. Guided inference of this form gives advantage to short regional models that are capable of being easily transformed to regression or classification tasks, depending on the child node (i.e., response variable). In these settings, BNs can be assessed at the unit level, ranked probability estimates can be used for various predictions, and the output can be used as a model fit indicator on a given system.

Probabilistic graphical models have been used in a range of different applications, both predictive and theoretical ones. The chapters 2 and 3 of the current work describe a PGM to question the frequency of experiencing dissonance, stemming from conflicting beliefs, and the subsequent disengagement of some of these beliefs as a means to resolve dissonance, within different political ideologies. We believe that using a probabilistic graphical model to answer theoretical questions such as the above, where the variables have intrinsic randomness and are difficult to define accurately, is more fitting than using a more traditional method, such as an agent-based model. Then, in chapter 4 we collect empirical data from Edinburgh, UK, and Napoli, Italy, to validate the conflict model and test its predictions.

### 1.1.6 From cognitive dissonance to polarisation

We use the results from the conflict model to explore how political agents choose to associate with others in order to minimise conflict. We argue that, on the one hand, people's confirmation bias (i.e., the tendency to look for information that confirms existing beliefs), and on the other hand, people's conflict avoidance tendency, drives individuals to form sub-groups usually consisting of like-minded others. Political polarisation is one of the outcomes of this process, which in turn leads to *echo chamber formation*.

### 1.1.7 Political polarisation and echo chambers

Polarisation is a wide concept that refers to the presence of distinct groups differing on one or more attributes. Political polarisation, in particular, refers to distinction based on political ideology. Within each polarised cluster, it is frequent to differentiate between mass polarisation, which takes place among the electorates, and elite polarisation which takes place among the elected. Here, I focus on mass polarisation. This is usually driven by narrow circles of like-minded peers, at the expense of a comprehensive multi perspective and evidence based understanding of public affairs; a phenomenon called *echo chambers*.

Literature on echo chambers derives from the theory of selective exposure, which poses that information consumers willingly select to be exposed to content which is congruent with their beliefs whilst avoiding uncongenial views (Sears & Freedman, 1967). Before the internet era –in which the accessible news sources were negligible– selective exposure in the pursuit of information did not generally emerge in circumstances of mass persuasion; however, post-internet people are able to reach (a huge amount of) information easier and modify what they want, thus they are more likely to select the content they will be exposed to (Tewksbury, 2005; Cass, 2007; Garrett, 2009). Online social networks have exacerbated this via their compound capability to enable users-news interaction in novel ways and to make use of complicated user-tracking algorithms to feed users with ideologically congenial content and drive ongoing engagement (Beam & Kosicki, 2014; Spohr, 2017).

Echo chambers are not purely an artefact of online social media. It has intensified this phenomenon between groups, but people have the motive to avoid opinion challenges (Mutz and Martin, 2001; Sunstein, 2001b; although see also Garrett, 2009 for an opposing suggestion) and endorse information supporting their own beliefs independent of social media popularity. To conclude, although online social networks might foster a fertile ground for the emergence of echo chambers, the actual causes have more to do with certain psychological factors.

One of the most widely used tools for investigating echo chambers formations has been agent-based modelling. More generally, agent-based modelling has been used to explore how simple rules on the micro level are able to generate emergent patterns on the macro level.

### 1.1.8 Agent-based modelling

There is a growing interest between social scientists in using system modelling tools to investigate how parts of a intricate phenomenon interact, persist or change, and ultimately understand the internal drivers of such systems. One such tool is agent-based modelling (ABM; e.g., Barrett et al., 2005; Yang et al., 2011). Agents (i.e., units) are granted characteristics and initial behaviour rules that direct their (inter) actions. Stochasticity can be added when defining agents' characteristics, when deciding which agents interact with one another, and how they gather information and take choices. The model is iterated to simulate temporal dynamics which allows for a distribution of feasible results for the given system. The micro entities called *agents*, are whatever changes their behavior responding to input from other entities (e.g., agents and/or environment; Auchincloss & Garcia, 2015). There are at least three benefits of using the agent-based technique over more conventional modelling approaches, like top-down methods of non-linear dynamics where associated state variables are aggregated (e.g., through a differential equation). The agent-based technique: 1) picks up emergent phenomena, 2) offers a natural-like environment for the investigation of specific mechanisms, and 3) is flexible, especially regarding the production of geo-spatial frameworks.

ABM is capable of accommodating high diversity in agent traits and interactions between units and environment, along with features such as dynamics, feedbacks, and adaptations, which are unfeasible to capture with conventional statistical model approaches (Macy & Willer, 2002; Auchincloss & Garcia, 2015). Units can be described at various levels, including agent or group level (e.g., political groups, organisations, etc.). Scientific investigations that entail substantial diversity within and across units and various geo-spatial and relational components are well suited to ABM (Grimm & Railsback, 2005). In social science literature, simulations are used to investigate dynamic phenomena having various populations and surroundings like legal and health services, town bodies, individual citizens, families and more. Some ABMs involve thorough data and strive for realism (Barrett et al., 2005) while others are quite abstract (Yang et al., 2011; Axelrod, 1997).

To implement interactions between units one can use geo-spatial networks, or a mixture of structures (as stressed in Alam & Geller, 2012). This would be much more complicated to describe by mathematics (see also Axtell, 2000). Importantly, ABMs can adjust units' behaviour on the basis of interactions at a given length and direction. Furthermore, ABMs offer a powerful and flexible framework for adjusting the intricacy of units (i.e., their behaviours, level of reasoning, aptitude to learn, capability to evolve, and ways of interacting with their surroundings). One more aspect of flexibility is the capability to adapt degrees of description and aggregation. It is convenient to work with aggregate units, sub-groups of units, or single units, with differing degrees of description concurring within a system. Therefore, ABMs are particularly useful when the right degree of description or complexity is not known and finding a fitting degree demands exploration.

One crucial limitation of ABM is the computational limit introduced by their complexity. By definition, ABMs regard systems at a dis-aggregated degree. This granularity involves the explanations of potentially several unit qualities and behaviours, and their interaction with the surroundings. An easy method to tackle this kind of problem in agent computing is via repeated iterations, methodically altering initial settings or parameters to evaluate the strength of the outcomes (Axtell, 2000). However, there is an upper boundary to the length of the parameter space that can be tested, and this procedure may be computationally expensive, and therefore, time-consuming. The large computational demands of ABMs remain a restriction when modelling massive systems (see Parry & Bithell, 2012), although this limitation shrinks as time goes by since computing power is increasing rapidly.

Although significant effort has been devoted to studying social networks and echo chamber formation, rigorous theoretical questions concerning morality, political orientation and group formation (i.e., echo chambers) is still lacking. In chapter 5 we use an ABM to investigate group formation between political agents who experience conflicting beliefs with one another. The inherit heterogeneity among our agents, and the interaction between agents –but also between agents and environment– render ABM a suitable computational approach for our study. Furthermore, to derive the initial parameters fed in the ABM we use the predictions of the conflict model presented in chapters 2 and 3.

### 1.1.9 From echo chambers to belief updating

The research of the emergent network qualities, like the distribution and density of the sub-groups, and its evolution through time, would be interesting in light of the effect of such time-varying topology on spreading dynamics. In addition, the emergence of sub-group structures where similar agents are relatively isolated from the rest of the sub-group is interesting also with respects to the recently explored emergence of online echo chambers, where misinformation spreads and persists (Garrett, 2009; Bessi et al., 2015; Del Vicario et al., 2016; Fränken & Pilditch, 2020). An interesting idea, which relates to the spread of misinformation in and out of echo chambers, is to explore how people update their beliefs when someone from their sub-group shares with them an opinion different from/congruent to theirs. We differentiate between a moral belief updating process, which we argue can rarely be broken down rationally, and non-moral belief updating which is usually rational. We use a Bayesian normative framework to simulate how a rational agent should update her beliefs.

### 1.1.10 Bayesian modelling

Even though the statistical instruments most widely employed by scientists in the discipline of psychology over the last century are frequentist, Bayesian methods are increasingly getting popular (e.g., Mandel, 2014; Etz & Vandekerckhove, 2018). We briefly discussed Bayesian modelling in the context of PGMs, but Bayes's theorem has a significant place in the discipline of sociocognitive science in numerous ways. In brief, Bayes's theorem is presently deployed within cognitive sciences with three primary ways (M. D. Lee, 2011). First, it is employed as a theoretical framework describing how the brain reaches inferences regarding the world. This approach rigorously acts as a theoretical account of human reasoning and behaviour. It might seem counter-intuitive, but in this approach, the data –sometimes– are still analysed using conventional, frequentist ways (e.g., Juslin et al., 2011; Kinoshita & Norris, 2012; Gopnik et al., 2015).

A second exercise of Bayesian approaches in cognitive sciences is via hierarchical modelling. This approach attempts to link models of psychological mechanisms to empirical data. Hierarchical modelling enables the scientist to develop a granular, firm model of the mechanisms that are supposed to be elements of human cognition, and to compare predictions of behaviour between model and real human conduct. If the framework is able to efficiently forecast the real data, we know something about the likely processes that generate the seen behaviour. The usage of these hierarchical Bayesian models is expanding since 2006, faithfully pursuing the trend of Bayes as a theoretical framework for cognition (Van De Schoot et al., 2017). Some recent examples of this line of work are M. D. Lee (2011) and Ferreira et al. (2012), and Scheibehenne and Studer (2014).

Finally, another usage of Bayesian approaches in sociocognitive discipline is the deployment of Bayesian statistics on empirical data. For instance, Andrews et al. (2009) deployed a Bayesian ANOVA to analyse several computational models –on the basis of cognitive theory– regarding how people understand semantic representations. Furthermore, this approach has been used to compare experimental conditions on word-generalisations using $t$ tests (Voorspoels et al., 2015), to test contrast effects in category learning using hierarchical Bayesian models and Bayes factors (Voorspoels et al., 2012), and to compare models of learning and adaptation using Bayes factor (Steyvers et al., 2009), just to mention a few examples. It is apparent that Bayes's theorem has a significant role in cognitive sciences, and its primary applications are gradually altering from theoretical frameworks to a practical way for estimating models.

In the current dissertation, we fall mostly under the first and second categories. That is, in chapter 6 we use a Bayesian model as a theoretical framework which aids on drawing predictions with regards to how people should *optimally* update their beliefs in light of new evidence. Then we collect empirical data and compare model predictions to observed data.

### 1.1.11 Dissertation layout

To reiterate, in the current dissertation we mostly use computational techniques as a tool to explore sociocognitive phenomena relevant to political ideologies. Specifically, in chapter 2 we define a *probabilistic graphical model* to investigate the frequency different political ideologies experience conflicting beliefs, and thus disengage from some of these beliefs as a means to resolve the conflict. Then, in chapter 3 we revise the conflict model and update two of its aspects: 1) we add interaction between agent's *moral foundations*, and, 2) add the concept of agreement between *event's moral foundations* (see chapter 3). We then go on to validate the model and test its predictions by collecting empirical data collected independently in Edinburgh and Napoli. In particular, in chapter 4, we test whether the predictions presented by the conflict model hold true also in empirical data. In chapter 5, we use the validated predictions of the conflict model, and feed them to an *agent-based model* that investigates echo chambers and political polarisation in different political ideologies. The agents follow only two rules: 1) they are attracted towards groups consisting of like-minded others, while, 2) they are repulsed from groups where they have a high chance to conflict with the rest members. Last, in chapter 6, we explore how individuals revise their moral beliefs in light of new evidence. We use a Bayesian model to draw normative predictions of a rational agent, and then we deploy two experiments to test the model predictions.

In order to make to make the content of each chapter *almost* independent (although not all of them can be independent since the results of some chapters directly influence the decisions taken/parameter used in some other chapters) we specify hypotheses and research question within each chapter. For the same end, each chapter has each own appendix section. https://www.overleaf.com/project/5ee7472f4bb4ed0001370f43

# Chapter 2

# Modelling the frequency of moral disengagement on the political continuum

## 2.1 Introduction

Conflicting political views and polarised beliefs are ubiquitous in both professional contexts and everyday life. For example, the recent death of George Floyd has induced heated political discourse on social media (ICantBreathe, 2020). Similarly, a potential decrease of global pollution observed during the COVID-19 pandemic has served as a battleground for conflicting political beliefs and hate speech (IFLScience, 2020). Other recent contexts in which conflicting political views clashed include discourse on holiday destinations (Traveller, 2020) or even preference debates over cats vs. dogs (LADbible, 2019).

A potential reason for the disputes arising from conflicting political beliefs is that political beliefs are ingrained in morality (Miles & Vaisey, 2015), and morality is by its very nature contradictory (e.g., de Burgh, 1930; Brand-Ballard, 2003). That is, morality can be a body of standards or principles derived from a code of conduct from a particular philosophy, religion, or culture, and between some moral propositions there might exist logical incongruity. This occurs when the propositions, taken together, yield two conclusions which form the logical, usually opposite, inversions of each other, which in turn could trigger a conflict between (and/or within) individuals. To illustrate the contradictory nature of morality, consider an individual who opposes abortion as they consider it as murder, while at the same time, they support the death penalty—a pattern which is mostly found in people belonging to the right wing of the political spectrum.

Alternatively, consider an individual who is doing their part to keep the environment clean with the overarching goal of making the world a better place for their children, yet they prefer using their car to work instead of catching the bus, –a moral contradiction that might be more frequent in leftist folks.

In the above examples, conflicting moral values result in a specific theoretical chain of reactions: when an individual has conflicting ideas or behaviours s/he tends to experience a feeling of mental discomfort called dissonance (Festinger, 1957, 1962; Brehm & Cohen, 1962) leading to an alteration in one of the attitudes, beliefs or behaviours to reduce the discomfort and restore balance (Bandura et al., 1996; Bandura, 1999, 2002). Since liberals and conservatives uphold different moral values (Graham et al., 2009; Haidt & Graham, 2007), it is reasonable to question whether the probability of moral conflict, is also different between different political orientations. To theoretically tackle this question, we build a Probabilistic Graphical Model (PGM) drawing on the results of the Moral Foundations Theory (MFT; Haidt & Joseph, 2004, 2007; Haidt, 2012), the theory of cognitive dissonance (Festinger, 1962, 1957), and the moral disengagement theory (Bandura, 1999).

### 2.1.1 Moral foundations theory and political orientation

Moral foundations theory (MFT; Haidt, 2001; Haidt & Joseph, 2004, 2007; Haidt, 2012) provides a well-defined framework of the roots and impact of moral intuitions. MFT adopts an evolutionary approach and describes morality functionally. That is, it regards the objective of morality is to ”suppress or regulate selfishness and make social life possible” and that this moral mechanisms consist of ”interlocking sets of values, practices, institutions and evolved mechanisms" (Haidt & Bjorklund, 2008, p. 70). According to MFT, humans posses some psychological mechanisms developed/evolved to produce acute, intuitive judgements. We feel those as gleams of affect that steer us towards moral (dis)approval (Haidt & Joseph, 2004). There are five mechanisms, or modules which have been identified

and thoroughly researched: [1] 1) Care/harm, 2) Fairness/cheating, 3) Loyalty/betrayal, 4) Authority/subversion, 5) Disgust/purity (see section 1.1.4 for a detailed summary of these foundations).

The former two foundations (care/harm and fairness/cheating) are jointly labelled *individualising* as they evolve around judgements relative to individuals. The other three foundations are labelled *binding* because they evolve around support and security of the team/group (Graham et al., 2011). Although all five are claimed to be universal to people, there is substantial variation –both on the individual and the group level– in how these foundations are cultivated and expressed. These modules are primarily regarded as evolved and innate traits (i.e., "organised in advance of experience"; Haidt & Joseph, 2011), but they can greatly altered by social interactions. Important role in reinforcing or retarding the growth of certain modules in a given person play the institutions and shared social values found in various cultures (Haidt, 2012; Haidt & Joseph, 2004). For instance, people who are raised in a society characterised by respect to tradition and expectations that group requirements precede individual needs are more probable to grow a greater faith on the binding modules when they make moral judgements. On the contrary, persons who grew up in an environment that promotes individual independence more than group needs is more probable to reinforce a greater support for individualising modules.

A significant component of MFT in developing these foundations is the culture and the surroundings of the given individual. Having said that, MFT also regards foundations as stable, dispositional, and genetically affected (Haidt, 2012). Specifically, MFT poses: "human beings have the five foundations as part of their evolved first draft, but ... there

[1] Although there is also evidence for a sixth foundation, Liberty/Oppression (Iyer et al., 2012), in the current dissertation we do not take it into account since the relevant research is scarce relatively to the main five foundations.

is heritable variation" (Graham et al., 2009, p. 1031).

Finding the drivers of political attitudes is an old task for sociologists, who pose that people usually adopt extreme political beliefs on the grounds of scarce information. Political preferences regularly appear to be more rationalised than rational –a phenomenon also observed with moral decisions. There is substantial controversy about the adjacent and distant causes of political ideology, but people mostly consent on that articulating a political view consists of forming a judgement of good versus bad, and judgements like this are frequently underlined by a moral character (Emler et al., 1983; Lakoff, 2002).

Moral foundations theory tries to explain why individuals have different ideologies, and why they adopt different political labels. Briefly, MFT proposes that person-level variation in political ideologies comes from systematic variation in how much people uphold each of the foundations. Liberals tend to uphold in a greater extent the individualising (i.e., care/harm and fairness/cheating) than the binding foundations (i.e., loyalty/betrayal, authority/subversion, and disgust/purity). Conservatives, on the other hand, value all five foundations more equally, although there is evidence to suggest that conservatives actually express the opposite pattern (i.e., value more the binding foundations; (Graham et al., 2009; Graham et al., 2011).

The relationships among moral foundations and political ideologies encourage treating MFT as a way to explain the drivers of ideology and political preferences, as well as they provide a sensible explanation of why these drivers can be genetically affected. However, there are three assumptions that have to be met in order to regard this theoretical explanation as valid. First, moral foundations have to be relatively persistent over time and dispositional. There is an overall consensus among evidence that political preferences are consistent over time and ideology is, partly, a dispositional characteristic (Ansolabehere et al., 2008; Krosnick and Alwin, 1989, but see also Converse, 1964). People can and do alter their views on specific subjects, or even change their political perspective all together, but largely, conservatives stay conservatives, and liberals liberals.

Second, alterations of foundations must systematically forecast alterations in political preferences. In fact, numerous researches directly make use of some variant of the statement *"moral foundations cause/explain/predict/shape political attitudes"* (e.g., Haidt, 2012; Inbar et al., 2012, 2009; Kertzer et al., 2014; Koleva & Haidt, 2012). Other studies are not so certain about the causal order; Graham et al. (2009, p. 1042), for instance, pose the question, "Do people first identify with the political left or right and then take on the necessary moral concerns, or do the moral concerns come first, or is there reciprocal influence or even an unidentified third variable the root of both?". This vagueness is echoed in studies that alternates moral foundations and political ideologies as predictor and outcome variables. Nonetheless, it is crucial to notice that in the current project we too stay agnostic about whether moral foundations are causes of political attitudes or the opposite.

The last assumption requires that these psychological modules (i.e., moral foundations) can be inherited. If moral foundations are evolved characteristics with a "heritable foundation", they have to, by definition be heritable (Graham et al., 2009; Haidt, 2012; Haidt et al., 2009). While it may not be possible to name individual genes connected to moral foundations, there is an obvious expectation that people grew up in alike environments with alike genetic structure must share alike moral values.

### Alternative accounts

Moral foundations theory has generated substantial research and discussion on the morality and political domains, but it has also received tangible criticism along the way. Two crucial points that have attracted heat by the Theory of Dyadic Morality (TDM) is 1) MFT's conclusion that liberals and conservatives have fundamentally different moral minds, 2) and MFT's broad definition of harm. TDM poses that liberals and conservatives essentially have the *same* moral mind. In other words, instead of proposing distinct and differentially activated mechanisms, TDM argues that moral judgement involves a common template grounded in perceived harm (the moral dyad; Gray, Waytz et al., 2012).

TDM emphasises the subjective nature of *harm* supporting that harm, like morality, is in the eye of the beholder. Indeed, a number of studies document the perception of harm in "harmless" cases of religious blasphemy, anti-patriotism, and aberrant sexuality (DeScioli et al., 2012; Kahan, 2007; Gray et al., 2014).

For example, let us consider a scenario outlined by anthropologist Fassin (2012): Oriya Hindu Brahmans believe it is exceptionally unethical for the older child to eat chicken just after their father's death. Westerner societies cannot see the wrong –or harmful– in this action, and they consider it as just a matter of religious protocol. Hindus, on the other hand, consider it the older child's duty to process the father's "death pollution" via a vegetarian diet. When the child eats chicken, they "place the father's spiritual transmigration in deep jeopardy" (Fassin, 2012, p.96). By understanding the perceived harm in these actions, even Western liberals can understand its perceived immorality.

MFT considers such perceived harm as mistaken, but dyadic morality sees these perceptions as legitimate. In the language of social anthropology, dyadic morality advocates for not only moral pluralism (accepting the legitimacy of different perceptions of morality) but also harm pluralism (accepting the legitimacy of different perceptions of harm). Harm pluralism suggests that different moral content such as purity and loyalty are (less prototypical) varieties of perceived harm. In contrast, MFT endorses harm monism, rejecting the legitimacy of harm in anything but direct physical or emotional suffering (Schein & Gray, 2015).

Taken together, we saw that moral foundations theory is not restricted to forecasting a relationship among moral foundations and political preferences. The causative means required for this relationship assumes that 1) moral foundations are consistent over time and dispositional (at least to the point that the credits on moral foundations at a given time point are predictive of the same credits at time point +1), 2) any significant alteration in moral foundations credits will have as a result alterations in political preferences, and that 3) moral foundations can be inherited. We also saw that MFT's definition of harm has been described as unnecessarily broad and that alternative accounts, such as TDM, are not so convinced that differences between liberals and conservatives can be explained by or attributed to deep differences in moral cognition.

### 2.1.2 Cognitive dissonance theory

The cognitive dissonance theory (Festinger, 1957; Festinger & Carlsmith, 1959; Festinger, 1962) posits that when a person upholds a number of *elements of knowledge* (i.e., beliefs/attitudes/behaviours; from now on *cognition*) which are pertinent but conflicting to each other, the person reaches a state of inconvenience; this obnoxious state was called "dissonance". The level of dissonance in connection to a cognition is $d/(d + c)$, where $d$ is the sum of cognitions which are conflicting with a specific cognition, and $c$ is the sum of cognitions agreeing with that same specific cognition (see Sakai, 1999; Shultz et al., 1999, for more well-defined mathematical models). Based on this theory, an individual is pushed by this discomforting state of being to decrease the dissonance they feel by reducing the inconsistency among the conflicting cognitions. A frequently evaluated mechanism for dissonance reduction is belief/behaviour alteration. For example, an individual who might argue for the freedom of individual expression in the arts but at the same time wants hateful speech to be regulated, might experience dissonance stemming from the contradiction between these two beliefs. One way to resolve this dissonance would be to disengage from, or *let go*, one of these beliefs. Extending cognitive dissonance theory, moral disengagement offers a number of mechanisms people use to disengage from their beliefs.

The theory of moral disengagement (MD) explains why individuals are capable of engaging in ruthless behaviour with no evident discomfort (Bandura, 1990a, 1990b, 1999, 2002). People who have higher scores in the MD scale are accustomed to the exercise of various cognitive strategies that prevents the self-regulation –that sociocognitive theory proposes control moral behaviour– by downplaying the ethical content of import of individuals' ruthless conduct.

MD theory proposes that disengagement acts over eight distinct cognitive systems which can be grouped in three functionally similar clusters. The first cluster consists of three of these systems (moral justifications, euphemistic labelling, and advantageous comparisons), and helps with the cognitive reframing of harmful behaviours to seem less ruthless –sometimes to the point that the even appear advantageous– to the person engaged with them. The second cluster deals with minimising the role of the person in the damage brought by their behaviour, and consist of two of these mechanisms: 1) displacement and 2) diffusion of responsibility. The difference between the two is that displacement refers to when an individual attributes the responsibility for their conduct to other individuals who usually directed –or they might have purposely overlooked– their behaviour (see also Kelman & Hamilton, 1989). On the other hand, diffusion refers to when an individual distributes the responsibility among the members of a team instead of any specific person. The rest three mechanisms (distortion of consequences, dehumanisation and attribution of blame) compose the last cluster which is responsible for minimising the consequences of one's behaviours or the assessment of discomfort these behaviours induce to other people. Contrary to the first cluster, this is not meant to restructure the action in a good way; instead, it acts by minimising the real impact the action have on other individuals.

All in all, these eight cognitive systems reframe the way people make –and experience– choices. Therefore, MD is a specific inclination to invoke cognitions that enable people to reframe their behaviours to seem less restless, minimise their role in the consequences of their behaviours, or mitigate the discomfort they induce to other individuals, thus disengaging the self-sanctions that sociocognitive theory proposes direct moral behaviour.

Although the empirical work on moral disengagement was initiated mostly for forecasting aggressive and antisocial behaviour of young kids and adults (e.g., Bandura et al., 1996; Bandura et al., 1991), a substantial effort has been directed towards moral disengagement and organisational corruption (e.g., C. Moore, 2008; C. Moore et al., 2012; Zhao et al., 2019).

Corruption is defined as immoral behaviour carried out to aid organisational profits, that might or might not be beneficial to the actors (Clinard & Quinney, 1973; Schrager & Short Jr, 1978; Szwajkowski, 1985). In the long run, immoral choices are sparsely within an organisation's greater good –they can be expensive and even endanger an organisation's continuation. But how could MD facilitate organisational corruption? To answer this question we first need to discuss how MD works.

As discussed in the previous paragraphs, MD aids people in escaping the self-sanction that follows acting unethically. People who score higher on MD are more probable to take deviant actions that are contra to other people interests (Detert et al., 2008; C. Moore et al., 2012), and conveniently use MD mechanisms to reframe or rationalize their immoral acts. This is backed from research which has specifically stated that MD has a significant part in the procedure of organizational corruption (C. Moore, 2008). MD mechanisms operate either over the redefining of the ruthless action to seem innocent or over defencing the diversions by referring to norms; for example ”everyone else is doing the same”. It is therefore apparent that MD facilitates corruption.

Taken together, these theories lead to quite a few testable hypotheses, one of which is with respects to the differences in MD frequency between different political ideologies. In particular, we saw that moral foundations theory (Haidt & Joseph, 2007; Haidt, 2012) bridges morality to politics by suggesting that different ideologies have different moral profiles, with liberals endorsing the individualising foundations more, while conservative uphold all foundations almost equally (Graham et al., 2009). Assuming that conservatives denser moral systems leads to having more moral beliefs (Turner-Zwinkels et al., 2020), it is likely that they also find themselves –more frequently than liberals– in situations where some of these beliefs contradict with each other. We also saw that the cognitive dissonance theory (Festinger & Carlsmith, 1959; Festinger, 1962) posits that people having contradictory cognitions (i.e., beliefs/behaviours) experience an unpleasant state of being called dissonance and have the need to get out of there by disengaging from one or more of their beliefs. Following this line, if conservatives do have more moral beliefs, they should morally disengage from them more frequently than liberals. We computationally test this hypothesis using a probabilistic framework.

### 2.1.3 Theoretical framework and relevant studies

The framework of probabilistic graphical models (PGMs) sets out an instrument for exploiting structure in complex distributions to depict them compactly, and in a way that enables them to be build and utilised efficiently (Koller & Friedman, 2009). To do so, a

PGM uses a graph-based representation over a potentially high-dimensional space, which eases representation by defining a skeleton for that space. Instead of encoding the probability of any feasible assignment to all of the variables in our domain, we can *split* the distribution into thinner *factors*, each over a much smaller scope of possibilities. That way allows us to specify the entire *joint* distribution as a product of these factors.[2] In the current study the most states of our theoretical model are quite abstract and thus include inherit uncertainty. In particular, we do not seek to examine the intrinsic mechanisms of internal conflict (i.e., how does conflict come about), rather we are interested in how frequently people experience conflict and resolve it, and we try explaining the frequency variance on the basis of an agent's morality. In other words, we are modelling the initiating causes, and a possible resolution strategy, of conflict, but not the psychological state of conflict.

Previous work has addressed modelling belief systems as networks –both political (e.g., Brandt et al., 2019) and moral (e.g., Turner-Zwinkels et al., 2020) networks– and there have been numerous detailed efforts to model the psychological state of cognitive dissonance/conflict itself (e.g., Shultz & Lepper, 1996; Van Overwalle & Jordens, 2002; Read & Monroe, 2007), but to our knowledge, there is no work directly linking moral belief systems to dissonance resolution.

## Belief systems as networks

A recent work (Brandt et al., 2019) used social network analysis (SNA) to explore if operating (i.e., attitudes on matters) or symbolical (i.e., emotive links to political parties and labels) elements are more essential in the larger system of the whole belief structure. Specifically, they modelled the interrelations of positions and beliefs regarding politics as a network of cooperating units (see also Boutyline & Vaisey, 2017, for a seminal implementation of this method). They treated every unit as a way to evaluate an operating or symbolic element of the belief structure which contained all the related units and their links with each other. Their results replicated research of many years (e.g., A. B. Cohen,

[2] A more detailed and technical overview of PGMs can be found in appendix A.2, but for now this brief introduction should be enough to allow comparisons to other studies.

2003; Bartels, 2000; Fiorina, 2002): symbolical elements were more important than operating elements to the general structure and were nearer to several forms of political belief systems. In other words, people's symbolical links to political groups and labels are a more central component than the real political attitudes. This replicates work in which less thorough methods were used for investigating the rich organisation of political beliefs systems (Converse, 1964; Kinder & Kalmoe, 2017; Malka & Lelkes, 2010). These results can also be linked to probabilistic accounts of belief polarisation, where two individuals with incongruent prior beliefs both reinforced their beliefs after coming across identical evidence (Jern et al., 2009; Jern et al., 2014; Cook & Lewandowsky, 2016)

Although SNA has a rich history it tends to break with huge databases. Because of the vastness of (online) social networks and intricate dependencies in such data, bogus finding rates are not adequately controlled, which renders the recognition of noteworthy signs and relations hard (Efron et al., 2007). However, the conjunction of SNA and PGMs has shown some promise in addressing such problems (Farasat et al., 2015).

Extending Brandt et al.'s (2019) work, a recent study (Turner-Zwinkels et al., 2020) used a probabilistic approach in conjunction with SNA to chart a subdivision of moral ideologies anticipated by MFT. In this study, ideologies are regarded as networks, with moral beliefs portrayed as units lined by straight relationships. This approach moves past latent variable ones (e.g., factor analysis) that regard moral beliefs as exchangeable pointers of a hidden factor, towards a non-static portrayal of moral beliefs, where nodes in the system are able to directly affect one another. This provides insights in the organisation of moral structures, by depicting how particular moral values are correlated. Exploiting three data-sets (2 from the U.S., 1 from New Zealand), network techniques were applied to validate MFT's findings. It was found that liberal's moral structure has a larger divergence between individualising and binding foundations, while this divergence was insignificant for conservative's moral structure. Although this study supports MFT's proposition that individualising and binding foundations are dissimilarly upheld by liberals and conservatives, there is at least one limitation regarding how should one interpret these results. The network models used in the study model (one-sided) relations, displaying bi-directional correlations among items, but cannot be use to deduce causation. For that, it does not speak to whether upholding particular MF items will be most probable to encourage a person to perceive themselves as liberal or conservative. Rather, their findings offer an explanation of group differences.

### Models of cognitive dissonance

There is a substantially larger amount of computational work on cognitive dissonance. The majority of these models describe attitude change as a cognitive consistency-seeking issue, and initial such efforts used a connectionist approach, relying on either constraint satisfaction or attributional theory. At their time, these efforts were among the first examples of a new field called computational social psychology. There are three main theoretical models in this area which have shown enormous potential for formal modelling of social cognition: the Consonance Model (Shultz & Lepper, 1996), the Adaptive Connectionist Model for Cognitive Dissonance (Van Overwalle & Jordens, 2002), and the Recurrent Neural Network Model for long-term attitude change following cognitive dissonance decrease (Read & Monroe, 2007).

### Consonance model of cognitive dissonance

The consonance model (Shultz & Lepper, 1992, 1996, 1998) represents particular cognitive structures (e.g., belief systems) that are presumed to contain incongruities between attitudes and behaviours, regarding a specific attitudinal item (e.g., either an attitude or behaviour). These structural incongruities trigger an unpleasant state (i.e., cognitive dissonance) that is handled based on a constraint satisfaction algorithm.

The units in the network reflect a person's cognitions (behaviours, attitudes, actions, affects, beliefs). Positive correlations among cognitions are depicted by an excitatory (i.e., positive) connection, while negative associations by an inhibitory (i.e., negative) link. The network undergoes numerous iterations of activation adjustments. In each iteration, the value of every node is adjusted in parallel, with the value of each node being informed based on the activation of the nodes to which it is linked either by an excitatory or an inhibitory link. In such a cognitive architecture, the problem of decreasing the dissonance is tackled as a congruency-seeking mission: iterations do not stop before the whole system has reached an equilibrium: the activation levels of each node is stable, that is, alterations in the activations between iterations are negligible. This settled state is called *consonance*. Formally[3], the consonance added by a given node $i$ is as follows:

$$consonance_i = \sum_j w_{ij} a_i a_j \tag{2.1}$$

where $w_{ij}$ is the weight between nodes $i$ and $j$, $a_i$ is the value of the recipient node $i$, and $a_j$ is the value of the sender $j$. The total consonance in the system is the tally of the values provided by the equation 2.1 across all recipient nodes in the system:

$$consonance_n = \sum_i \sum_j w_{ij} a_i a_j \tag{2.2}$$

.

This model adopts a localist approach: each notion corresponds to a node. The cognitive system consists of cognitions –where each cognition is represented as a node (unit)– and connections among the cognitions –where each connection is depicted as a path among the units–. The units can be either active or inactive; if inactive, their value is 0, otherwise, if active, its value can vary continuously up to 1. Activation is affected by connection weights: stronger weights have a greater influence on each unit's activation value.

Each belief (unit) is also characterised by its immunity to change: this parameter allows updating beliefs with a modifying activation value. The resistance value is a multiplier from 0 to 1, which gets multiplied by the activation value. The greater the magnitude of the immunity parameter, the weaker the immunity to change. Since units can get diverse activations through their weights, their immunity to change is a significant parameter of the whole consonance of the system.

A particular characteristic of this model is that the beliefs are represented as nodes having two smaller units with opposing activations. That is, one smaller unit might have a positive value and represents one pole of the cognition, while the other might

[3] For a thorough technical description Shultz and Lepper (see 1996).

have a negative value and represents the other pole. This kind of representations is described as a *dimension*. Dimensions go on a range from positive values to negative values. Conveniently, this indicates whether any two beliefs (nodes) are dissonant with each other: if they resulting weight of their link is negative they are dissonant, otherwise they are consonant.

This also implies that the consonance or dissonance of the units depends on the sign of the connections between any two nodes: two nodes (beliefs) are consonant if the weight of their connection is positive.

The model has successfully simulated behaviour in several paradigms of insufficient justification where individuals are more probable behave in a way which conflicts their beliefs when they are presented with a shorter as opposed to a greater reward (e.g., Shultz & Lepper, 1992, 1996, 1998; Shultz et al., 1999). However, a critical limitation of the model is that it only pits two beliefs against each other at a time. The theory of cognitive dissonance (Festinger, 1957; Festinger & Carlsmith, 1959) states that dissonance can be triggered by two or more conflicting cognitions, and although the consonance model sums up the dissonance coming from all pairwise connections of the beliefs, it does not take into account potential interactions that might exist in one's belief system. In other words, although the consonance model can model dissonance as a function of the relative overall strengths of two competing beliefs, it does not take into account the structure of the system these beliefs are embedded. Furthermore, the consonance model lacks a representation of affect, which is one of the upgrades Van Overwalle and Jordens's (2002) implemented on their adaptive connectionist model.

### Adaptive connectionist model of cognitive dissonance

Contrary to the consonance model (Shultz & Lepper, 1996) which relies upon the notion of equilibrium, the adaptive connectionist model of cognitive dissonance (AC-CD; Van Overwalle & Jordens, 2002) relies on the concepts of causality and causal explanation. In this model, causal explanations are sought in the cognitive system (knowledge, experience) in order to provide justification for dissonance depletion. Dissonance depletion relies upon behavioural alteration which, in turn, is based upon a causal explanation of the discrepancy between expected and real outcomes (Van Overwalle & Jordens, 2002).

The AC-CD framework has approaches the cognitive dissonance in an attributional manner which posits that when the conflicting behaviour is ascribed to an individual's responsibility, that individual is pressured to alter their attitude. On the contrary, when an external demand (e.g., compensation or threat by an authority figure) provides adequate justifications for engaging in the discrepant conduct, the pressure to reduce dissonance is lower. The model implements a distributed representation of cognitions (i.e., attitudinal objects: attitude, behaviour, affect) or external factors (threats) represented by multiple nodes. Cognitive dissonance is the distance between expected and real outcomes[4] –in other words, the difference between the affect and the behaviour. This is a somewhat different definition of dissonance from that articulated by Cooper and Fazio (1984).

The model is a typical feed-forward network with activation values following a specific route from input to output. This has intended theoretical implications: by virtue of being unidirectional, the connections depict causative descriptions of the results, and the relevant weights reflect magnitude of causative effect or magnitude of behaviour. In the AC-CD model, dissonance reduction relies on an attitude change mechanism implemented via a learning algorithm intended to minimise the distance between expected and actual results (actions and affect). Once the causative interpretations of these differences are identified, they supply the justifications for behavioural alteration, which reduce dissonance via minimising the mismatch between expectation and experience.

The representation of dissonant cognition, behaviours, and affect relies on the *Affect-Behaviour-Cognitive* system of attitudes (Rosenberg et al., 1960). Attitudes are relationships –stored as memories– between beliefs regarding the attitudinal item and two kinds of reactions: 1) behaviour toward the attitude item and 2) affective results of the interplay with the attitudinal item. The weights related to the links in the network reflect the magnitude of the attitude; they start from zero and can go either way. The nodes in the

[4] This is similar to how the activation value gets updated in the Temporal Difference (TD) learning algorithm, where the difference (error) between the expected ($t + 1$) and the current ($t$) reward is calculated and taken into account when updating.

network reflect causes and results. The causes incorporate the attitudinal item and the external factors. The results comprise the behaviour and affect. Results can have various causes.

When comparing the aforementioned models with comparative analysis –and using the same dissonance paradigm as in (Van Overwalle & Jordens, 2002)–, the AC-CD has higher performance which can be ascribed to its learning features (Van Overwalle & Jordens, 2002). However, this model comes with its own limitations. The Delta-learning algorithm used in the AC-CD primarily adapts the weights in a way that is capturing instructor's own estimates alterations regarding the attitude item, that is, the instructor *guides* the system how the estimate is altered. The solution to this issue comes to give the recurrent neural network framework for long term behavioural change (Read & Monroe, 2007). This model captures the estimate alterations using a back-propagation learning algorithm able to learn estimates and alterations in estimates, and thus, give insights for long term attitude alteration.

### The recurrent neural network for long-term attitude change

The recurrent neural network (RNN) for attitude change resulting from dissonance reduction was first described by Read and Monroe (2007). The RNN model leverages a combination of Gestalt constraint satisfaction, similar to the consonance model, coupled with a back-propagation learning algorithm that drives long-term attitude change. Thus, it is a type of hybrid between the consonance and AC-CD model approaches.

Here, a node can represents a belief, an attitude, a behavioural outcome/alternative, an evaluative task, an object, and an instrumental/contextual factor (i.e., compensation in Festinger and Carlsmith's (1959) paradigm of counter-attitudinal advocacy). The cognitive structure is rather flexible, allowing for adjustments of beliefs/attitudes via back-propagated learning, as the behavioural outcomes may or may not be those expected. The network is initialised with evaluations of objects and later on adjusts those evaluations based on new evidence which follows the assessments of behavioural alternatives. This assessment adopts a constraint-satisfaction approach.

The constraint satisfaction paradigm strives to achieve consistency among differing assessments of the same item. The differences in the assessments follow from the comparison between expectations and real results of various behavioural alternatives. After a number of iterations, the system can distinguish old from new evaluations –after the learning process has been established making the model capable of distinguishing behaviours counter to its previous experience–, and to modify the dissimilarities between evaluations as a task to reduce dissonance.

The connections (i.e., weights) between the nodes of the network take values in the range of -1 to 1 allowing for a way of reciprocal restraint among evaluative units. This technique was also used in the consonance model (Shultz & Lepper, 1996) where it simply denoted a bi-polar kind of assessment. In the RNN model, however, this mechanism reflects the possibility of separate assessments (positive/negative) of the attitude item, enabling the network to distinguish between behavioural alternatives.

While the Delta-learning algorithm in the AC-CD mostly adapts the weights in order to pick-up instructor's own estimate alterations regarding the attitude item, the Contrastive Hebbian Learning (CHL) back-propagation algorithm deployed in the RNN model can pick-up the estimate alterations on its own, which is a major improvement. In other words, in the AC-CD model, the researcher *tells* the system how the estimate changed, while in the RNN, the system can detect evaluation alterations, triggering attitude alteration by adapting the cognitive structure itself (a reflective and flexible feature). Adjustments of the cognitive structure then decrease dissonance in a balanced or consistency seeking way (Read & Monroe, 2007).

Although these connectionist models differ in the structure of the networks and the choice of paradigm of dissonance reduction they implement, they nevertheless share a common focus on formally exploring how dissonance can be reduced within an attitude/belief structure, primarily via changes to the strength of attitudes or beliefs. However, these are all process models for dissonance reduction in a generic sense. They are agnostic regarding the cause(s) of conflict between cognitions and experiences that drive dissonance itself. Similarly, they do not systematically explore how individuals might differ in their responses to dissonance as a function of either their belief structures or reaction to dissonance.

The model we present in the current project is more specific than the aforementioned attempts. First, when exploring dissonance drivers it takes into account the causes of conflict between cognitions and/or beliefs –bringing into play the moral belief structure of the individuals. Furthermore, our model also addresses how political agents might differ in how frequent they experience conflict, and thus dissonance, as a function of their ideology. More broadly, our model extends pre-existing computational attempts in dissonance literature, by shifting the focus from exploring the intrinsic mechanisms of dissonance to exploring its frequency of occurrence within individuals. In other words, in our model, although there is a formal implementation of dissonance, we mostly regard it as a black box while focusing more on its prevalence.

### 2.1.4 Current study

While the connectionist models described above have provided significant insights into the possible processes involved in cognitive dissonance, they are limited in ways not suitable for our central question. An orthogonal approach to connectionist networks, particularly When confronted by a problem where the variables have intrinsic uncertainty and/or the data are scarce or incomplete, is probabilistic graphical modeling (M. J. Johnson et al., 2016). One is not better than the other for all purposes, and neural networks and probabilistic graphical models (PGMs) have been fruitfully combined (see for example M. J. Johnson et al., 2016; Siddharth et al., 2017).

Here we introduce a PGM which shows the usefulness of jointly modelling abstract moral foundations/values with political ideologies for predicting the frequency of conflict. Investigating how frequently people experience conflict in a political world, and how statistical dependencies in the moral belief structure of the political agent contribute to the frequency of the conflict are thus important for the socio-cognitive literature, as well as politics and morality research, with implication for, for example, the study of echo chambers (e.g., Bikhchandani et al., 1992; Watts, 2002; Whalen et al., 2018; Madsen et al., 2018), polarisation of political ideologies (e.g., Leifeld, 2014) or advocacy organisations' attempts to shape public debate (e.g., Bail, 2016)

We chose to use as a tool a PGM over a neural network as the most states of our theoretical model are quite abstract and thus include inherit uncertainty. Our computational problem can be defined as trying to find the probability of a categorical event (i.e., having discrete levels) given the probability of a particular state of the world. More specifically, we are trying to model the probability of conflict —and as a result the probability of a strategy to resolve conflict (e.g., moral disengagement)— given a snapshot of a person's moral values and a given state of the world (i.e., a moral scenario where each of the five foundations might or might not be triggered). In other words, in the real world we cannot simultaneously satisfy all of our moral values at every moment they might be salient. Thus, we want to formally model what happens when an individual's moral values intersect a world/event that does not necessarily respect those values. The theoretical question we are trying to answer is whether there are any differences between various political orientations in the frequency of moral disengagement strategies used when there exist conflicting beliefs. Moral foundations theory (Haidt & Joseph, 2004, 2007; Haidt & Graham, 2007; Graham et al., 2009) helps on bridging the gap between political orientation and conflicting beliefs, by linking the five moral foundations to different political positions on the left-right continuum.

We expect our model to identify differences in the frequency of conflict –as a result disengagement– between different political orientations. In particular, since, according to MFT (Haidt & Joseph, 2004, 2007; Haidt & Graham, 2007; Graham et al., 2009), liberals value more two out of five moral foundations while conservatives tend to credit all of them almost equally, we expect that conservatives will experience conflicting beliefs more frequently. Our model makes five core assumptions: 1) all foundations generate the same number/amount of beliefs; 2) the frequency with which the (simulated) world's events involve beliefs associated with any particular foundation is equal; 3) the probability of two beliefs to conflict (with respect to the state of the simulated world), when everything else is kept constant is the same across and within moral foundations (this assumption will be relaxed in a later version of our model 3); and 4) the probability of detecting conflict (i.e. individual sensitivity) is equal for all individuals; and 5) the probability of morally disengaging as a dissonance reduction strategy is equal across all (simulated) individuals. Therefore, given these starting assumptions, conservatives should more frequently experience conflict than liberals as a direct function of their relatively greater number of equivalently valued moral foundations (Graham et al., 2009; Haidt, 2012), and thus, they should exhibit moral disengagement at higher frequency.

## 2.2 Methodology

### 2.2.1 Theoretical description

This section assumes some prior knowledge of probability theory, information theory, and graph theory. Appendix A.2 discusses the foundations and key concepts of PGMs, but for a more detailed review the interested reader should read a more extensive textbook, such as Koller and Friedman (2009).

The nodes in our graph $\mathcal{G}$ (see graph 2.1) represent discrete random variables. The nodes are connected to each other via links or edges which imply a causal relationship between the nodes, with the direction of the causation flowing towards the direction of the edge.

#### Political orientation space

For the sake of explanation, let us divide the model into three theoretical spaces: the political orientation space which includes the person's moral foundations, the conflict space, and the events space. The *political orientation* space includes a random variable for political orientation, along with the five moral foundations all of which are represented as random variables in our model. Note that we do not differentiate the moral foundations on the individual level, rather, we only draw a distinction between the individualising and the binding clusters of foundations. The probability distribution of the political orientation random variable spans over the political spectrum, where on the one end we have extreme liberals and on the other, extreme conservatives. The political orientation variable is the parent of the five moral foundations of the individual.

As we have seen from MFT (Graham et al., 2009; Haidt, 2012), people's endorsement of the foundations is closely related to their political ideology. Specifically, liberals strongly endorse the individualising foundations (harm and fairness), whereas the further to the right one's ideology, the more probable that one values the binding foundations to a greater extent. Therefore, in the model we condition the five moral foundations on the political orientation of the virtual agent. The first two nodes make up the individualising foundations while the remaining three make up the binding foundations. The levels of each of the moral foundation variables represent the probability of it being active or inactive, which differs as a function of political orientation. That is, there is a *conditional probability distribution* (CPD) over the levels of the moral foundations, and it is conditioned on the different levels of the political orientation variable. For example, the probability of *fairness* being active is higher if the agent's score on the political spectrum is towards the left side. On the other hand, if we observe a conservative agent, the probability of any of the *binding foundations* being active is greater.

### Event space

The structure introduced so far is inadequate to capture contradictory beliefs, since in its current state it only describes if the moral foundations are enabled conditioned on the agent's political orientation. That is, it only describes if an agent has particular moral values, not whether those values coincide or conflict in a given instance. To instantiate the latter dynamic, we introduce the event space. This describes the moral facets of an external situation faced by the agent in question. It includes the five moral foundations, each of which can, but need not, be active aspects of a given situation, and thus ”push” the agent's internal foundations in a given direction.

### Conflict space

The *conflict space* reflects standard dissonance theory, in the sense that incompatible cognitions/moral values generate dissonance. The conflict space consists of the conflict variable, which is the child of both the five moral foundations of the agent and the five moral foundations of the event (we talk about the event space below). As a reminder, the cognitive dissonance theory states that two or more cognitions can be dissonant if the obverse, or opposite, of one cognition follows from the other. Put simply, if two cognitions are enabled but point towards opposing actions/conclusions then they clash with each other. This is exactly what the underlying function of the conflict variable tries to capture, but in probabilistic terms. The more probable the existence of two or more contradictory beliefs, the more likely the agent will experience conflict which will, in turn, trigger dissonance, and potentially, disengagement.

Conflicting beliefs/cognitions will place the agent in an unpleasant state of dissonance which the agent will feel the need to decrease. As reviewed in subsection 2.1.2, one strategy to lessen the burden of the dissonance is to disengage from one or a few of the dissonant beliefs. Moral disengagement theory (Bandura, 1999) introduces eight such strategies that allow individuals to justify their dissonant beliefs/behaviours, thereby reducing the aversive motivational state of dissonance. For reasons of simplicity, in the current model we conceptually 'average over' the eight mechanisms and we end up with a binary-valued variable of the sort *O/I* indicating whether or not disengagement occurred. We have, in a sense, hard-coded that the more probable the conflict, the more likely the agent will choose to disengage. We consider disengagement as being more of a choice (square node in the graph 2.1 than a random variable (circle) to distinguish it from conflict which is more like an involuntary incident.

Having the above handy, let us visit an example. Let us consider the *authority* and *ingroup loyalty* foundations, which underlie virtues of leadership and followership, and virtues of patriotism and self-sacrifice for the group, respectively. Let us also consider an individual who went to vote on the morning of 3/11/2020 in the US, and encounter armed federal agents sent by Donald Trump to intervene in ballot-counting efforts (Lichtblau, 2020). Under most nationalistic-patriotic ideologies, not voting is distinctly unpatriotic, which conflicts with the individual's high value on patriotism and ingroup loyalty. At the same time, the individual is *pushed* to comply with the rules and laws of the country of which they are a part, consonant with upholding obedience to authority. In either way, a part of this individual has to walk out disappointed, after the conflict he/she experiences. The frequency of scenarios like this we are trying to model.

Coming back to the conflict variable after having fleshed out what the event foundations are, it is easier to describe how a conflict comes about according to our model. So far we have seen that for a conflict to exist we need (at least) two enabled foundations which point in opposing directions, but this is not all it takes. The conflict function also takes into account how probable, or in this setting, how *intense* two or more foundations are uphold by an agent. If only one of the two foundations is highly valued, then the agent will disengage from the less active foundation, and thus, will result in a non-conflicting end. Returning to our food company example, if the individual does not care too much about the rules and policy of their company, then they will most probable not experience conflict, as one of their foundations will be almost deactivated. To conclude, a conflict requires both *highly* and *comparably* active foundations, or else the agent will disengage from, or drop one of their conflicting beliefs.

### 2.2.2 Formal description

We now turn our attention to describing the model formally. We start by formally introducing the nodes and edges of the model, and the cardinalities of the random variables. Then, we move on to explaining the function for the conflict variable and describing the algorithms we used to make inferences. Last, we describe the moral disengagement decision node and how the agent *decides* whether or not to disengage by maximising the utility of the given choices.

As we have seen, the round nodes in our graph $\mathcal{G}$ represent random variables each having a number of levels. A discrete probability distribution is assigned to the levels of each random variable. We aim to represent a joint distribution $P$ over a set of random

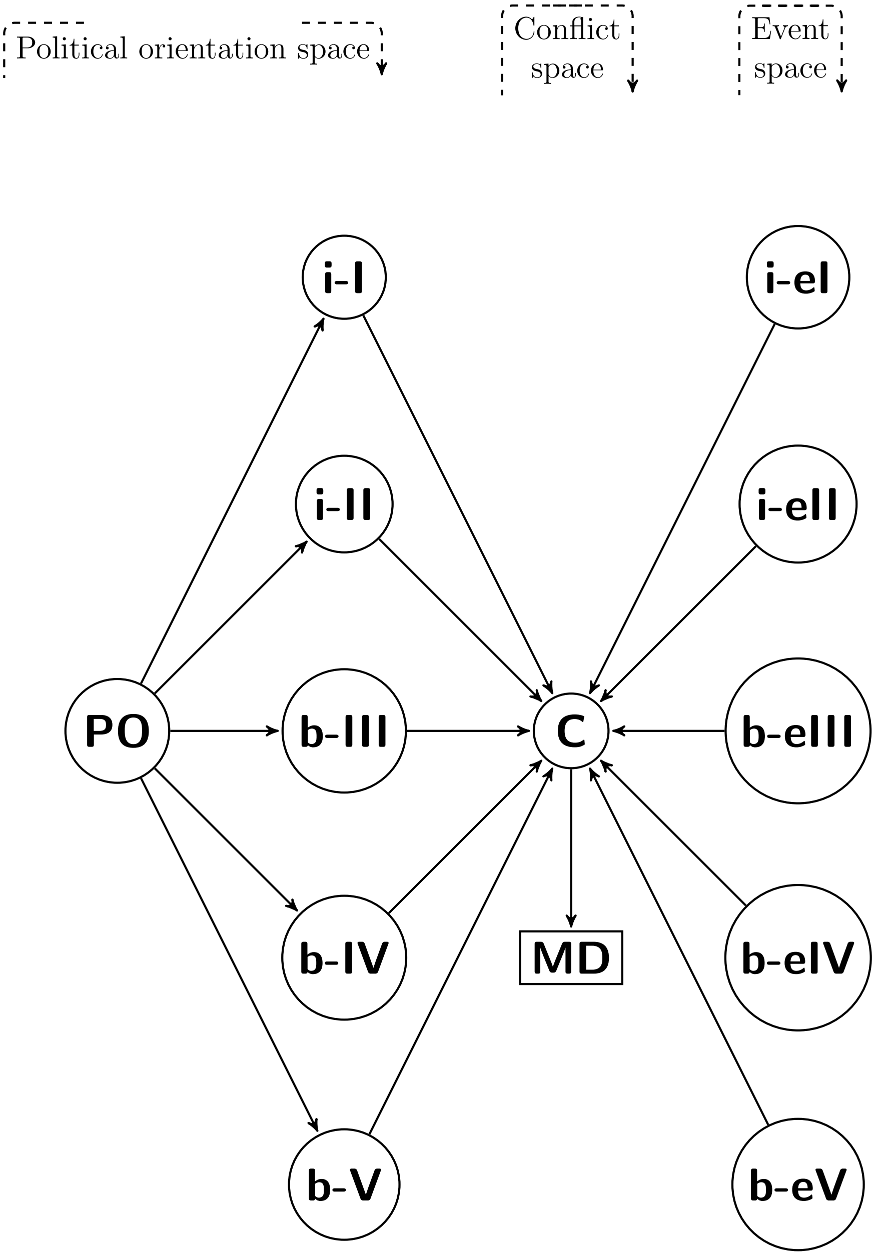


*Figure 2.1.* The DAG $\mathcal{G}$

variables $\mathcal{X} = \{X_1, ..., X_n\}$. The set of random variables in our model consists of the following: $\{PO, < agent'sMFs >, < event'sMFs >, C\}$, where $PO$ stands for political orientation, $C$ for conflict, and $MF$s are the five moral foundations. Note that, for the sake of space and simplicity, we sometimes use $< agent'sMFs >$ to refer to the agent's moral foundations, e.g. $I_a, II_a, III_a, IV_a, V_a$. Similarly, we refer to clusters of foundations, binding or individualising (refer to relevant section in intro chapter here), for a given agent (a) as $MF_a^b$ or $MF_a^i$, respectively. Note, also, that the disengagement node is missing from the set since it is not a random variable but a decision node as mentioned in subsection 2.2.1.

The variable $PO$ can take five values $po^1$, $po^2$, $po^3$, $po^4$, $po^5$ representing the political orientation of the agent spanning from very liberal to very conservative with *neutral* in the middle level. Next, each of the agent's five $MF$s ($MF_a$) can take on one of four levels $mf_a^1$, $mf_a^2$, $mf_a^3$, $mf_a^4$, from disabled to very likely to be enabled. Note that $MF_a$s are conditioned to the $PO$ variable, which results in $5 \times 4 = 20$ possible $MF_a$ states, for each foundation. The event's $MF$ ($MF_e$) consist of five levels, $mf_e^1$, $mf_e^2$, $mf_e^3$, $mf_e^4$, $mf_e^5$, representing that each event can trigger either a *left* or a *right* reaction (see alse subsection 2.2.1). Note here that left and right have nothing to do with political orientation. They just denote opposing directions, meaning that they could as well be south and east, or yes and no. Last, the conflict $C$ random variable consists of just two levels $c^0$ and $c^1$, representing that the agent either does or does not experience conflict. The conflict variable's parents are all of the $MF_a$s and $MF_e$s, hence making up a high dimensional space.

With respect to the function which defines the conflict variable, we theorise conflict as being more probable when two or more beliefs, which are *comparably* and *highly* valued, clash with one another. Formally, we have:

$$P(C \mid MF_a, MF_e) = 1 - |e| * d \tag{2.3}$$

where $e$ is the *energy*, or how enabled all foundations are, and $d$ is the total difference between the foundations. Energy, $e$, is calculated as:

$$e = \frac{\hat{f}}{\max f * 0.5} \tag{2.4}$$

where $\hat{f}$ is the mean of the left, $l$, and right, $r$, sums of the agent's foundations multiplied by the event's foundations resulting in pointing either to the *left*, or the *right*[5]:

$$\hat{f} = \frac{l + r}{2} \tag{2.5}$$

We scale $\hat{f}$ with the constant $\max f$ which is the maximum value possible out of the sum of $l$ and $r$. This constant, of course, depends on the given parameters; in our case $\hat{f} = 15$. We halve this value as the maximum *mean* of the *left* and *right* sums cannot exceed $\max f/2$. We give more details on the parameters in the 2.2.3 subsection.

The difference $d$ in equation 2.3 is computed by calculating the difference between the left and the right activation values and dividing this by $\max f$:

$$d = \frac{l - r}{\max f} \tag{2.6}$$

We do not have to halve $\max f$ here, since, given the initial parameters, the value of the difference can exceed $\max f/2$ but not $\max f$.

Now we turn to explaining how we make inferences with our model. We can extract useful information from a model by querying its joint distribution. One of the most common query types, and the one used here, is the *conditional probability query* see appendix A.2. Formally, conditional probability rules requires the following:

$$P(Y \mid Ev = ev) = \frac{P(Y, ev)}{P(ev)} \tag{2.7}$$

We can calculate every instance of the numerator $P(Y, ev)$ by marginalising out the entries in the distribution that match to assignments consistent with $Y, ev$. In particular, if $W = \mathcal{X} - Y - Ev$ are not either query or evidence random variables, then:

$$P(Y, ev) = \sum_{w} P(y, ev, w) \tag{2.8}$$

[5] Note again that left and right refer here to (mis)alignment of moral foundations, both in the agent's psychology and in the event itself, not political ideology.

Every term $P(y, ev, w)$ in the summation is an entry in the joint distribution, since we are talking about all of the network variables: $Y, Ev, W$.

The denominator of the conditional probability equation 2.7 $P(e)$ can be simply computed as follows:

$$P(ev) = \sum_{y} P(y, ev) \tag{2.9}$$

allowing us to stash the calculation of equation 2.8. By calculating both equation 2.8 and equation 2.9, we are able to divide each $P(y, ev)$ by $P(ev)$, to retrieve the wanted conditional distribution $P(y \mid ev)$. Notice that this procedure matches with taking the vector of the marginal probabilities $P(y^1, ev), ..., (P(y^k ev)$ (where $k = |Val(Y)|$) and *renormalising* the entries to total to 1.

Sometimes, calculating the whole distribution of a given graph, and then summing out the terms we want to sum out is unpractical or even unfeasible if we are talking about a large distribution. This approach is not good enough as it takes us back to the exponential explosion of the joint distribution issue that PGMs were precisely designed to prevent. For that reason, we use a particular set of algorithms which help us avoid this blowup, but they still give us an exact solution (there are also approximation algorithms, but our distribution does not require such techniques). As it happens, the graphical structure that enables a compact representation of complex probability distributions also help loosen up the inference problem. Specifically, we can use dynamic programming strategies to perform inference even for certain large and complex distributions in an acceptable amount of time. Two helpful ideas are:

- Exploiting the structure of a Bayesian framework which allows for a scarce amount of dependencies between the joint distributions and the random variables

- We could calculate the results once and cached them, so to avoid generating them multiple times.

The main actions in the variable elimination algorithm explained below can be seen as a manipulation of factors. Specifically, the essential operation that is performed when calculating the distribution of some subset of variables is that of marginalising, or summing out variables. Formally, if we have a distribution over some variables $\mathcal{X}$, and we want to calculate the marginal of that distribution over a subset $X$, then we can view this distribution as an action on a factor. Thus, if $X$ is a set of variables, $Y$ a variable which does not belong to the set $X$, $Y \notin X$, and $\phi(X, Y)$ is a factor, then we define the *factor marginalisation* of $Y$ in $\phi$, denoted $\sum_Y \phi$, to be a factor $\psi$ over $X$ such that:

$$\psi(X) = \sum_{Y} \phi(X, Y) \tag{2.10}$$

This operation is also called *summing out* $Y$ in $\psi$. The take away message from this operation is that we only sum up entries in the table where the values of $X$ match up.

Before presenting the algorithm, we need first to lay out a few important concepts. An essential notice used in implementing inference in graphical models is that the operations of factor product and summation behave exactly the same as do regular product and summation operations. In particular, both operations are *commutative*, so that $\phi_1 \times \phi_2 = \phi_2 \times \phi_1 =$ and $\Sigma_X \Sigma_Y \phi = \Sigma_Y \Sigma_X \phi$. Products are also associative, so that $(\phi_1 \times \phi_2) \times \phi_3 = \phi_1 \times (\phi_2 \times \phi_3)$. Most importantly, there is a simple rule that allows us to exchange summation and product: If $X \notin Scope[\phi_1]$, then

$$\sum_{X} (\phi_1 \times \phi_2) = \phi_1 \times \sum_{X} \phi 2. \tag{2.11}$$

Having laid out these properties, we are now ready to introduce the variable elimination algorithm. Let us demonstrate the procedure by actually applying the algorithm on a simpler version of the conflict model. Consider the network depicted in figure 2.2, which is a subgraph of our model. The chain rule for this network asserts that:

$$\begin{aligned} P(I, II, C, eI, eII) &= P(PO)P(I)P(II)P(eI)P(eII)P(C \mid I, II, eI, eII) \\ &= \phi_I(I)\phi_I I(II)\phi_e I(eI)\phi_e II(eII)\phi_C(I, II, eI, eII) \end{aligned}$$

We will now apply the variable elimination algorithm to compute $P(C)$. We will use the elimination ordering: $I, II, eI, eII$.

1. Eliminating $I$: We compute the factors

$$\begin{aligned} \psi_1(C, I, II, eI, eII) &= \phi_I(I) \cdot \phi_C(C, I, II, eI, eII) \\ \tau_1(C, II, eI, eII) &= \sum_{I} \psi_1 \end{aligned}$$

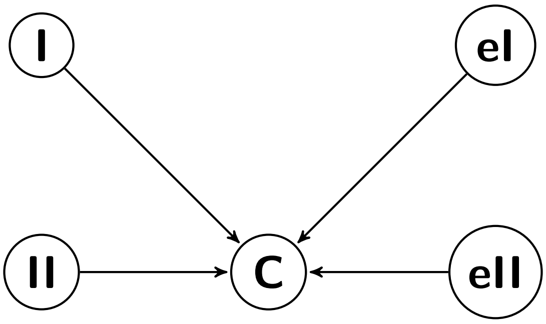


*Figure 2.2.* Subgraph of the main model having removed PO and MD and 3-5 MFT nodes.

2. Eliminating $II$: Note that we have already eliminated one of the original factors that involve $C$ – $\phi_C(C, I, II, eI, eII) = P(C \mid I, II, eI, eII)$, but we introduced the factor $\tau_1(C, II, eI, eII)$ which involves $C$. Hence, we now calculate:

$$\psi_2(C, II, eI, eII) = \phi_{II}(II) \cdot \tau_1(C, II, eI, eII)$$
$$\tau_2(C, eI, eII) = \sum_{II} \psi_2$$

3. Eliminating $eI$: We compute the factors

$$\psi_3(C, eI, eII) = \phi_{eI}(eI) \cdot \tau_2(C, eI, eII)$$
$$\tau_3(C, eII) = \sum_{eI} \psi_3$$

4. Eliminating $eII$: We compute the factors

$$\psi_4(C, eII) = \phi_{eII}(eII) \cdot \tau_3(C, eII)$$
$$\tau_4(C) = \sum_{eII} \psi_4$$

And there we have $\tau_4(C) = P(C)$. We summarise these steps in table 2.1 Note that we could have used any elimination ordering, and in this example it might not be apparent,

Table 2.1
Variable elimination steps over $P(C)$.

| Step | Variable eliminated | Factors used | Variables involved | New Factor |
|---|---|---|---|---|
| 1 | I | $\phi_I(I), \phi_C(C, I, II, eI, eII)$ | $C, I$ | $\tau_1(C, II, eI, eII)$ |
| 2 | II | $\phi_{II}(II), \tau_1(C, II, eI, eII)$ | $C, II$ | $\tau_2(C, eI, eII)$ |
| 3 | eI | $\phi_{eI}(eI), \tau_2(C, eI, eII)$ | $C, eI$ | $\tau_3(C, eII)$ |
| 4 | eII | $\phi_{eII}(eII), \tau_3(C, eII)$ | $C, eII$ | $\tau_4(C)$ |

but other orderings might introduce factors with much larger scope, unnecessarily slowing down performance[6]. A generic form of this algorithm can be found in Algorithm 1.

**Algorithm 1** Sum-Product Variable Elimination Algorithm

```
procedure SUM-PRODUCT-VE(
$\Phi$, \\Set of factors
$Z$, \\Set of variables to be eliminated
$\prec$, \\Ordering on Z
)
    Let $Z_1, ..., Z_k$ be an ordering of Z such that $Z_i \prec Z_j$ if and only if $i < j$
    for $i = 1, ..., k$ do
        $\Phi \leftarrow$ Sum-Product-Eliminate-Var($\Phi, Z_i$)
    end for
    $\phi* \leftarrow \prod_{\phi \in \Phi} \phi$
    return $\phi^*$
end procedure

procedure SUM-PRODUCT-ELIMINATE-VAR(
$\Phi$, \\Set of factors
$Z$, \\Variable to be eliminated
)
    $\Phi' \leftarrow \{\phi \in \Phi : Z \in Scope[\phi]\}$
    $\Phi'' \leftarrow \Phi - \Phi'$
    $\psi \leftarrow \prod_{\phi \in \Phi'} \phi$
    $\tau \leftarrow \sum_{Z} \psi$
    return $\Phi'' \cup \{\tau\}$
end procedure
```

### Incorporating evidence

Although algorithm 1 is useful for querying our distribution, it does not cover the cases where we want to calculate the probability of a given variable after having observed some evidence. For example, assume we observe the value $i^1$ (the first moral foundation is enabled) and $ei^{-1}$ (the first event's moral foundation is pointing to the left). Now our goal is to calculate $P(C \mid i^1, ei{-}1)$. From this result, we can calculate the conditional probability as described in equation 2.7, by renormalising the probability of the evidence $P(i^1, ei^{-1})$. But how do we compute $P(C, i^1, ei^{-1})$?

[6] In our implementations we also use an algorithm which finds the least computationally expensive variable elimination order. However, we are not going to describe this algorithm here as we believe its explanation is irrelevant with the purposes of this project.

The answer depends on the fact that an unnormalised measure derived from introducing evidence into a Bayesian network is equivalent to a *Gibbs distribution.* So we can now view this computation as summing out all of the entries in the *reduced factor*: $P[i^1 ei^{-1}]$ whose scope is $\{C, I, eI\}$. This factor is no longer renormalised, but it is still a valid factor.

According to this statement, we can now use the sum-product variable elimination algorithm to calculate $P(Y, e)$. We implement the algorithm to the factors in the network reduced by $E = e$, and eliminate the variables in $\mathcal{X} - Y - E$. The resulting factor $\phi^*(Y)$ is then $P(Y, e)$. To get $P(Y \mid e)$ we have to renormalise $\phi^*(Y)$ by multiplying it by $\frac{1}{\alpha}$ to get a legal distribution (i.e., a distribution that conforms to the probability rules), where $\alpha$ is the total over the entries in our unnormalised probability space, which represents the probability of the evidence. We summarise this process of computing conditional probabilities in our Bayesian conflict model in Algorithm 2.

**Algorithm 2** Sum-Product Variable Elimination Algorithm for computing conditional probabilities

```
procedure COND-PROB-VE(
$\mathcal{K}$, \\A network over $\mathcal{X}$
$Y$, \\Set of query variables
$E = e$, \\Evidence
)
    $\Phi \leftarrow$ Factors parameterising $\mathcal{K}$
    Replace each $\phi \in \Phi$ by $\phi[E = e]$
    Select an elimination ordering $\prec$
    $Z \leftarrow= \mathcal{X} - \mathcal{Y} - \mathcal{E}$
    $\phi^* \leftarrow$ Sum-Product-VE $(\Phi, \prec, Z)$
    $\alpha \leftarrow \sum_{y \in Val(Y)} \phi^*(y)$
    return $\alpha, \phi^*$
end procedure
```

We now move from the random variables of our model to the moral disengagement *decision* node. The task of these two types of nodes are different in the sense that, on the one hand, random variables represent information processing under uncertainty, while decision nodes represent deciding how to act in the world on the basis of that information processing. For instance, assuming moral disengagement as one option among several (e.g. revise one's moral beliefs), each option might lead to one of many outcomes, which the agent can prefer to differing degrees.

In the conflict model we use a non-trivial case of decision setting, where the result of each action is not completely deterministic. In this case, we must take into account both the probabilities of various results and the preferences of the agent between these results. Thus, here it is inefficient to define a preferred ordering on the various results. On the contrary, we must be able to attribute preferences to complex scenarios involving probability distributions over possible results. For example, considering a scenario where the agent has conflicting beliefs and where disengaging could reduce some of the burden conflict imposes, the agent will not just deterministically choose to morally disengage from one of their beliefs; the probability of conflicting beliefs will be taken into account and thus a distribution will be formed around disengaging or not. The *decision theory* framework lays out a formal base for this kind of reasoning. According to this framework, we should assign numerical *utilities* to the diverse possible outcomes, encoding the agent's preferences. Then the agent would have to *maximise the expected utility*, which is the foundation for rational decision making under uncertainty (Levin, 2006; Briggs, 2014).

We now move on to formally describing the basic decision-making task or our model and we introduce the principle of maximum expected utility. To do so, we use a simplified version of our model. Before introducing our simplified example, let us lay out some terminology. We use the word *profit* to describe the positive outcome (i.e., relief) of disengaging from conflicting beliefs. We assume that disengaging from non-conflicting beliefs is something the agent disfavours. Now, consider a political agent who finds herself in the following dilemma. She can choose to disengage from a potentially conflicting belief $A$, where the profit[7] of disengaging from that belief is 40 with probability .20, and 0 profit with probability .80. On the other hand, she can choose to disengage from a belief which is more likely to be conflicting $B$, so the profit of disengaging is 30 with probability .25 and 0 profit with probability .75. In other words, in the $B$ scenario, disengaging is less profitable (in absolute terms) but also less risky. In order for the agent to decide from

[7] The numbers assigned to the profit have merely an ordinal character, in the sense that 10 is greater than 5, and 80 less than 100, but other than that, they does not attribute any other meaning to the respective case.

which scenario is more profitable to disengage, she must compare her preferences between the two scenarios, each of which comes with a probability distribution over outcomes: the first scenario, denoted with $\pi_A$, can be written as $[40profit : 0.2; 0profit : 0.8]$; the second scenario, denoted $\pi_B$, has the form $[30profit : 0.25; 0profit : 0.75]$.

In order to decide which of these scenarios the agent prefers, it is not enough to just compare the values of the profits of each scenario, and say that she prefers 40 over 30 over 0. She also needs to take into account the probabilities associated with each of them. To do so, we use the concept of *utility*, where a higher utility value associated with an outcome indicates that this outcome is more preferred. We need to note here that utility values not only indicate an ordinal preference ranking between outcomes but they also carry information about their relative importance, that is, the relative values of different states tells us the strength of our preferences between them. This allows us to combine the utility values of different states, which in turn enables us to ascribe an *expected utility* to scenarios where we are not sure about the result of an action. Thus, we can compare two possible actions using their expected utility.

Formally, our decision-making node is defined by the following elements:

- a set of outcomes $MD\{md_0, md_1\}$, for no disengagement and disengagement;
- a set of possible *actions* that the agent can take $A = \{a_1, ..., a_4\}$ (these would be: *disengage when conflict*; *disengage when no conflict*; *do not disengage when conflict*; *do not disengage when no conflict*);
- a probabilistic outcome model $P : A \to \Delta_O$, which defines a lottery $\pi_a$, which in turn specifies a probability distribution over outcomes given that the action $a$ was taken;
- a *utility function* $U : O \to \mathbb{R}$, where $U(o)$ is the agent's preferences for the outcome $o$.

Here we need to note that the definition of an outcome can also incorporate the action taken. In other words, outcomes that involve one action $a$ would then get probability 0 in the lottery induced by another action $a'$.

With these, we can define the principle of *maximum expected utility* (MEU principle), which asserts that, in a decision-making situation $D$, we should choose the action $a$ that maximises the expected utility:

$$EU[D[a]] = \sum_{o \in O} \pi_a(o) \cup (o) \tag{2.12}$$

To make the formal definition of utility more concrete, consider a moral disengagement decision $I_F$ where a political agent is trying to decide whether to disengage from potentially conflicting beliefs. Although slightly unintuitive, for simplicity's sake in giving an example of utility computation, assume the agent is not aware if any of his beliefs conflict or not; thus, he has an arbitrary distribution representing his uncertainty: the conflict is either absent $c^0$, low $c^1$, or high $c^2$, with probabilities 0.5, 0.3, 0.2 respectively. The agent's profit, if he chooses to disengage, depends on the actual situation. If indeed the conflict is absent, the profit to disengagement would be negative since disengaging without conflict would require (potentially troublesome) cognitive effort to no purpose (outcome $o_1$); if conflict exists but is low, the profit would be small (outcome $o_2$); if it is high, the profit would be large (outcome $o_3$), if he does not disengage at all, outcome $o_0$, he earns nothing but loses something as we assume that dissonance is a negative motivational state (i.e., a necessary bad in some cases). We will now assign a value to the agent's utility for each of the four outcomes, like so: $U(o_0) = -2; U(o_1) = -7; U(o_2) = 5; U(o_3) = 20$. The agent's expected utility for the action of disengaging, $f^1$, is:

$$\begin{aligned} EU[D[f^1]] &= 0.5 \cdot (-7) + 0.3 \cdot 5 + 0.2 \cdot 20 \\ &= 2 \end{aligned}$$

His expected utility for the action of not disengaging, $f^0$ is 0. The action choice maximising the expected utility is therefore $f^1$.

Having fleshed out the theoretical and formal details of our model, let us now move in to its initial setup.

### 2.2.3 Initial setup

By hold to Laplace postulate 200 years ago ”When nothing is known about $X$ in advance, let the prior $p(x)$ be a uniform distribution, that is, let all possible outcomes of $X$ have the same probability”, we initialised the model using a uniform distribution for the political orientation and the event’s moral foundations, while we hardcoded the weights of CPD of the agent’s moral foundations. The CPDs of the agent’s moral foundation have the form of the figure 1.1, but instead of having a continuum political orientation variable, the models includes a categorical variable. The exact form of these CPDs can be found in appendix A.3, tables A.1 and A.2. Note here that we do not differentiate *within* political orientation clusterings (*individualising* and *binding* foundations), rather, we assume that individualising foundations (care and fairness) share the same values between them, as well as binding foundations (loyalty, authority, sanctity). For the conflict random variable, we used a theory driven function which assigns a probability distribution across all possible states of the model

## 2.3 Results

Before running the main simulations, we performed a few sanity checks to see if the results of the model seem plausible. This is actually a sort of validation step. Besides debugging the code and the algorithms for potential (logical) errors, we also run a few sanity-check simulations and tested them against existing results in the literature, or our theory-driven expectations.

### 2.3.1 Sanity checks

Based on previous research (Haidt & Joseph, 2007; Haidt & Graham, 2007; Graham et al., 2009; Haidt, 2012), we expected a conservative to value almost equally all five moral foundations, i.e. an almost uniform distribution over the agent’s moral foundations after we have observed a conservative. Formally: $P(I_a, II_a, III_a, IV_a, V_a \mid PO = conservative) \sim \mathcal{U}(a, b)$ or, in short, $P(MF_a \mid conservative) \sim \mathcal{U}(a, b)$. Running a diagnostic reasoning query, where the *flow* of reasoning goes from the parent to the child, we get the distributions depicted in Table 2.2, where we can see that we have similar distributions within the different *activeness* levels. The row we are more interested in is, however, the third row depicting the case where a given moral foundation is active at a medium level. Based on Graham et al.'s (2009) (see also Figure 1.1), conservatives should value all moral foundations almost equally, but not as high as liberals value the individualising foundations, which is captured by the third row of the table. All the five distributions are have a high (although not peaked) probability.

Table 2.2
Probability distributions over agent's moral foundation when a conservative has been observed.

| | | **Moral Foundations** | | | | |
|---|---|---|---|---|---|---|
| | | I | II | III | IV | V |
| **activeness** | disabled | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 |
| | low | 0.4 | 0.4 | 0.2 | 0.2 | 0.2 |
| | medium | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 |
| | high | 0.1 | 0.1 | 0.4 | 0.4 | 0.4 |

*Note.*This table is best read row-wise, where we can see that for a very conservative the binding foundations (columns 1-2) are highly likely to be lowly enabled, while the individualising foundations are highly likely to be highly enabled.

Let us now see if our model captures the subtle difference in the moral foundations activeness between moderate and extreme conservatives. As can be seen from Figure 1.1, although conservatives value all five moral foundations almost equally, extreme conservatives seem to uphold loyalty, authority and sanctity more than care and fairness. In other words, they value the binding foundations more than the individualising foundations, which is the exact opposite pattern we observe on extreme liberals. However, the magnitude of difference between the valuation of the binding and individualising foundations for extreme liberals seems to be larger than that of extreme conservatives. Let us now see if our model picks up this subtle but important difference. Comparing the distributions over the moral foundations of an extreme conservative $P(MF_a \mid v.conservative)$, versus an extreme liberal $P(MF_a \mid v.liberal)$, demonstrates the expected difference. Focusing on the differences between probability for high activation and probability for disabled status for binding (3-5) and individualising (1-2) foundations, extreme conservatives are more

Table 2.3
Probability distributions over agent's moral foundation when a very conservative has been observed.

| | | **Moral Foundations** | | | | |
|---|---|---|---|---|---|---|
| | | I | II | III | IV | V |
| **activeness** | disabled | 0.3 | 0.3 | 0.1 | 0.1 | 0.1 |
| | low | 0.4 | 0.4 | 0.1 | 0.1 | 0.1 |
| | medium | 0.2 | 0.2 | 0.4 | 0.4 | 0.4 |
| | high | 0.1 | 0.1 | 0.4 | 0.4 | 0.4 |

*Note.*This table is best read row-wise, where we can see that all four distribution are similar, with that being more apparent in the third row where all values are 0.3, that is, it is 30 percent likely to observe a given foundation in a conservative on the medium level.

likely to strongly value the binding foundations (row 4, columns 3-5), and weakly value the individualising foundations (row 2, columns 1-2, table 2.3). By contrast, extreme liberals follow almost the opposite pattern: highly valuing the individualising foundations (row 4, columns 1-2) while the last three foundations are more likely to be disabled (row 1, columns 3-5, table 2.4). Note that the difference between binding and individualising foundations is more intense for extreme liberals, that is, it is more probable for them to have highly activated the individualising foundations and disabled the binding foundations, while for extreme conservatives it is more probable to have highly activated the binding foundations but weakly activated the individualising ones (liberals: individualising being highly activated: 0.4, binding being disabled: .4; so the two extremes are *high* and *disabled*; conservatives: individualising being lowly activated 0.4; binding being highly activated: .4, so the two extremes are *high* and *low.* Specifically, the difference resides in columns 1 and 2, where extreme liberals have 0.1, 0.2, 0.3, 0.4, from *disabled* to *high*, respectively, while extreme conservatives have 0.1, 0.2, **0.4**, **0.3**, from *high* to *disabled*, respectively).

We see from the results that the model translates the point estimates taken from Graham et al., 2009 in a probability distribution fairly decently. Of course, these results do not add anything to our knowledge; they merely validate our computational instantiation of extant theory and its relevant data. That is a frequent procedure in translating a verbal theory into a computational model, as the details one has to decide upon when doing so

Table 2.4
Probability distributions over agent's moral foundation when a very liberal has been observed.

| | | Moral Foundations | | | | |
|---|---|---|---|---|---|---|
| | | I | II | III | IV | V |
| **activeness** | disabled | 0.1 | 0.1 | 0.4 | 0.4 | 0.4 |
| | low | 0.2 | 0.2 | 0.4 | 0.4 | 0.4 |
| | medium | 0.3 | 0.3 | 0.1 | 0.1 | 0.1 |
| | high | 0.4 | 0.4 | 0.1 | 0.1 | 0.1 |

*Note.*This table is best read row-wise, where we can see that for a very liberal the binding foundations (columns 1-2) are highly likely to be enabled, while the individualising foundations are highly likely to be either disabled or lowly enabled.

are essentially countless (see Oberauer, 2018, for an in-depth description of translating a verbal model to a theory). Building a computational model forces us to be explicit about the nuts and bolts of the cognitive mechanisms that we talk about in our theories

could Let us now move on to answering more interesting queries.

### 2.3.2 Querying the distribution

Our main question is whether there is any difference in the conflict and moral disengagement frequencies between different political orientations. We approach this via diagnostic reasoning, having first observed some evidence. Comparing just the probability of a conflict between conservatives and liberals we have $P(c^1 \mid liberal) \neq P(c^1 \mid conservative)$, and in particular $P(c^1 \mid liberal) < P(c^1 \mid conservative)$. For just this computation we will demonstrate the steps we took to calculate the probability of conflict one by one, but for the sake of simplicity we will not demonstrate these steps for the forthcoming computations as they involve almost the same equations. Formally, our problem is to compute $P(c^1 \mid liberal)$, but we can reduce it to computing the unormalised distribution

$P(C, liberal)$:

$$\begin{aligned} P(C, lib, < MF_a >, < MF_e >) = P(< MF_a >| \, lib)P(< MF_e >) \\ P(C \, |< MF_a >, < MF_e >) \\ = \phi_{<MF_a>}(< MF_a >, lib, C)\phi_{<MF_e>}(< MF_e >, C) \\ \phi_C(C, < MF_a >, < MF_e >) \end{aligned}$$

We then apply the variable elimination algorithm to compute $P(C, lib, < MF_a >, < MF_e >)$. We will use the elimination ordering $< MF_e >, < MF_a >, C$, as our algorithm suggests:

1. Eliminating $< MF_e >$: we compute the factors:

   $$\begin{aligned} \psi_1(C, < MF_a >, < MF_e >) = \phi_{<MF_e>}(< MF_e >, C)\phi_C(C, < MF_a >, < MF_e >) \\ \tau_1(C, < MF_a >| \, liberal) = \sum_{<MF_e>} \psi_1 \end{aligned}$$

   Note that we perform this step five times in total: each for the the five moral foundations of the event.

2. Eliminating $< MF_a >$:

   $$\begin{aligned} \psi_2(C, < MF_a >) = \phi_{<MF_a>}(< MF_a >, C)\tau_1(C, < MF_a >| \, liberal) \\ \tau_2(C \mid liberal) = \sum_{<MF_a>} \psi_2 \end{aligned}$$

   Note that we perform this step five times in total: each for the the five moral foundations of the agents.

   So we can now go ahead and compute $\tau_2(C \mid liberal)$ using equation 2.7:

3. Computing the conditional probability $P(C \mid liberal)$:

   $$P(C \mid liberal) = \frac{P(C \mid liberal)}{liberal}$$

.

Implementing these steps, we get the Table 2.5. As we can see, when we hold everything else equal, liberals' probability of experiencing a conflict is 0.21 while conservatives' probability of conflict is 0.28. It is important to note here that none of the event's moral foundations have been observed. To remind you, the event's moral foundations follow a uniform distribution for the four levels of being activated (activated low and pointing to the left, activated high and pointing to the left, activated low and pointing to the right, activated high and pointing to the right), while there is a small chance of not being activated at all. That is, when we have a political agent in potentially conflicting settings, then if this political agent is liberal, conflict is slightly less probable than if she was conservative. But so far we are only half-way through answering our question, which concerns moral disengagement and not conflict per se.

Table 2.5
Probability distributions over $P(C \mid liberal)$ and $P(C \mid conservative)$

| | | **Political orientation** | |
|---|---|---|---|
| | | Liberal | Conservative |
| **Conflict** | Disabled | 0.79 | 0.72 |
| | Enabled | 0.21 | 0.28 |

*Note.* We can see that liberals are less prone to conflict than conservatives, keeping everything else equal.

As a reminder, a decision making unit merely reflects the *willingness* of an agent to take an action. In our terms, the value the moral disengagement of an agent represents whether or not the agent is willing to disengage or not, under a given scenario. The exact value does not carry any other information other than the relative order and magnitude of willingness. Calculating the expected utility of moral disengagement actions between liberals and conservatives for the aforementioned scenario, we find that liberals' expected utility for disengaging after a potential conflict is 0.66, while conservatives' score is 0.99. That is, averaging across equally probable event types, conservatives are predicted to have greater expected utility for disengagement than liberals.

Moving on to more extreme political ideologies, further to the right and to the left of the spectrum where we have very liberal and very conservative agents. Repeating the same steps, we observe that extreme liberals' probability of experiencing a conflict is 0.21 while extreme conservatives' is 0.27 (see table 2.6). These results are quite interesting. First we can see that the model predicts extreme liberals will experience conflict less frequently than extreme conservatives in a scenario where the event's moral foundations

Table 2.6
Probability distributions over $P(C \mid V.liberal)$ and $P(C \mid V.conservative)$

| | | **Political orientation** | |
|---|---|---|---|
| | | V. Liberal | V. Conservative |
| **Conflict** | Disabled | 0.79 | 0.73 |
| | Enabled | 0.21 | 0.27 |

*Note.* We can see that very liberals are less prone to conflict than very conservatives, keeping everything else equal.

are unknown, but what is more interesting is the relative relationship of this trend to the trend depicted in the conservatives versus liberals table 2.5. Comparing moderate liberals to extreme liberals we observe that these two ideologies are almost equally probable to experience conflict, while extreme conservatives are slightly less probable to experience conflict than moderate conservatives. The upward trend of conflict from left to right breaks at conservatives, where extreme conservatives display a lower chance of experiencing conflict than conservatives. Table 2.7 depicts the differences between these two groups of political ideologies (extreme and less extreme).

Table 2.7
Difference between probability distributions over $P(C \mid extreme*)$ and $P(C \mid standard**)$

| | | **Political orientation** | |
|---|---|---|---|
| | | Very liberal - Liberal | Very conservative - Conservative |
| **Conflict** | Disabled | 0.0 | 0.01 |
| | Enabled | 0.0 | -0.01 |

*Note.* **Extreme* represent the two extreme political ideologies: very liberals and very conservatives; *Standard*** represent the two less extreme political ideologies: liberals and conservatives. We can see that very conservatives' and very liberals probability of conflict is less than this of conservatives' and liberals' probability of conflict respectively.

Computing the expected utility of the moral disengagement decision node for very liberal and very conservative agents, we find an expected utility for disengaging after a potential conflict of 0.63 for extreme liberals, while extreme conservatives' score is 0.97. We can see that there is a similar trend between extreme and less extreme ideologies with regards to conflict. That is, extreme liberals are less prone to disengagement than moderate liberals, given that their probability of experiencing conflict is less than that of moderate liberals. Similarly, extreme conservatives are less prone to disengagement, as well as to experience conflict, compared to moderate conservatives (see figures 2.3 and 2.4).

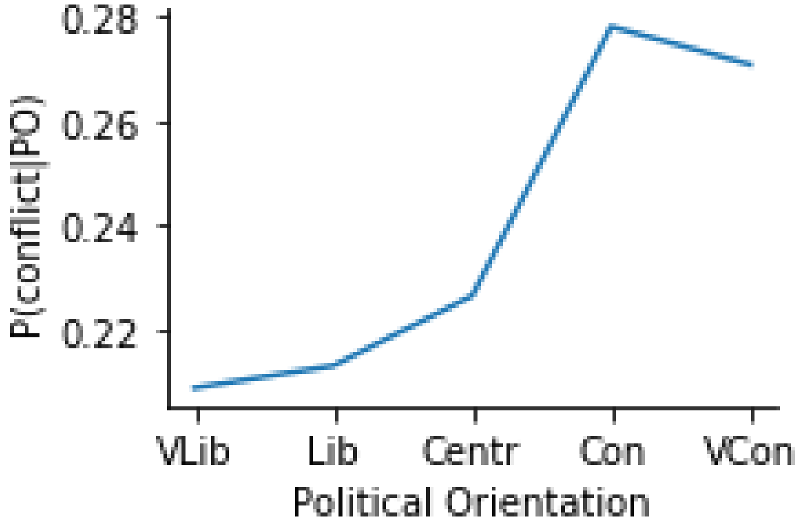


*Figure 2.3.* Probability of conflict over political orientation.

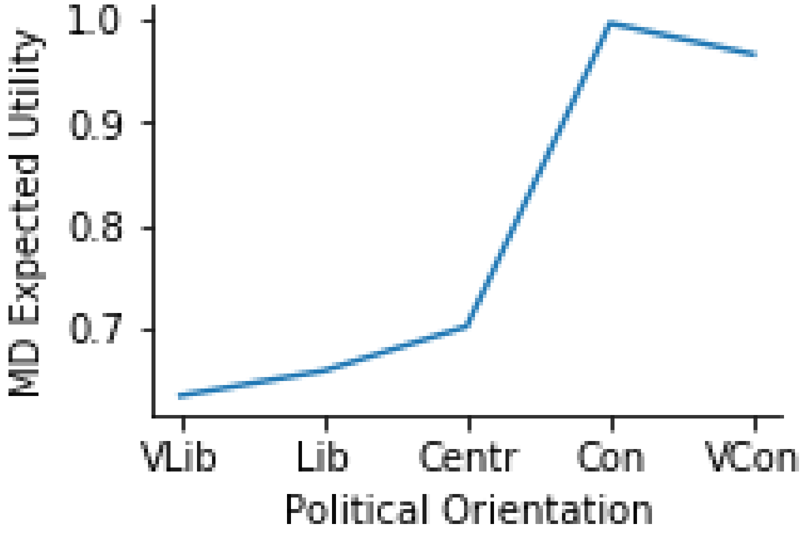


*Figure 2.4.* Expected utility of moral disengagement under different conflicting scenarios over political orientation.

The main questions of our project were answered by the aforementioned queries. These queries did not involve the event's moral foundations, in the sense that event's moral foundations were not observed. That means that in a given scenario they could have been either enabled or disabled. However, PGMs are very capable of responding to various queries without re-parameterisation. We now turn our attention to queries involving the event's moral foundations. Such queries can answer questions of the sort "what is the probability of conflict for a given political orientation when the event invokes *multiple binding foundation concerns*". To formalise this, we would observe, e.g., a liberal, and then a conflicting scenario which involves only binding foundations. To do so we could set two binding foundations as active in the event space, but 'pointing' in opposing directions.

There are three different types of conflicting scenarios we can set up for an agent: within individualising conflict (WIC), within binding conflict (WBC), and between individualising and binding conflict (BIBC). The WIC setting is a scenario where the two individualising foundations are both enabled but pointing to different directions, while the three binding foundations are all disabled. Formally, our problem is to find, for each political orientation, the probability of conflict given a within individualising conflicting event:

$$P(c^1 \mid I_e = Left^2, II_e = Right^2, III_e = 0, IV_e = 0, V_e = 0, PO) \tag{2.13}$$

where $Left^2$ and $Right^2$ indicate that the given foundation is enabled and pointing either to the *left* or to the *right*. $Left^1$ would denote an lowly activated foundation pointing to the left while $Left^2$ would denote a highly activated foundation pointing to the left. For the sake of simplicity, we will note mention the name of the foundation if all five foundations are observed. For instance, equation 2.13 is equivalent to:

$$P(c^1 \mid MF_e[Left^2, Right^2, 0, 0, 0], PO)$$

where $MF_e[Left^2, Right^2, 0, 0, 0]$ are the values assigned to the five event's moral foundations going from left ($I$) to right ($V$).

In terms of computations, the steps we follow to solve this query are the same as before, with the only difference being that now we have observed more evidence (we know the event's moral foundations). Running the model with these settings, we obtain the probability of conflict under a WIC scenario for each of the moral foundations: 0.22, 0.21, 0.17, 0.13, 0.09 for very liberal, liberal, centre, conservative, and very conservative agents, respectively (see also table 2.8, row 1). We can see that there is a downward trend spanning from very liberal to very conservative, with extreme liberals having the highest relative probability to experience conflict while extreme conservatives the least. That is of course to be expected, as liberals are the political agents who most value the individualising foundations (Graham et al., 2009), and since the conflict function requires highly held beliefs, liberals would be more likely to experience conflict. Regarding moral disengagement, we see that the same pattern arises, with expected utility values being higher for liberals and lower for conservatives: 0.66, 0.6, 0.42, 0.28, 0.20 for very liberals, liberals, centre, conservatives and very conservatives, respectively (see also table 2.8, row 2) in a scenario where only the individualising foundations are enabled.

Table 2.8
Probabilities of conflict, and expected utilities for moral disengagement for a within individualising conflicting scenario

| | **Political orientation** | | | | |
|---|---|---|---|---|---|
| | Very liberal | Liberal | Centre | Conservative | Very Conservative |
| **Conflict** | 0.22 | 0.21 | 0.17 | 0.13 | 0.09 |
| **MD*** | 0.66 | 0.6 | 0.42 | 0.28 | 0.20 |

**Moral Disengagement.* There is a downward trend from left (very liberal) to right (very conservative) for both probability of conflict and moral disengagement expected utility.

Naturally, the next query is for the probability of conflict in a WBC scenario $P(c^1 \mid MF_e[0, 0, Right^2, Left^2, Left^2], PO)$. Note here that there are more options for conflicting scenarios since the binding foundations are more numerous than the individualising foundations. For example, another WBC scenario could be $MF_e[0, 0, Left^2, Left^2, Right^2]$ or even with one of the three foundations being disabled $MF_e[0, 0, Right^2, Left^2, 0]$. The results are quite straightforward here too: a downward trend spanning from very conservative to very liberal. With these settings, we get the probability of conflict under a WBC scenario for each of the moral foundations: 0.11, 0.12, 0.17, 0.30, 0.31, for very liberal, liberal, centre, conservative, and very conservative agents, respectively (see also table 2.9, row 1) in a scenario where only the binding foundations are enabled. As before, the moral disengagement expected utility follows the same pattern as the probability of conflict: 0.25, 0.31, 0.47, 1.06, 1.14 (see also table 2.9, row 2)

So far, We have seen that liberals are more prone to experience conflict within an individualising conflicting event, and that conservatives are more prone to experience conflict

Table 2.9
Probabilities of conflict, and expected utilities for moral disengagement for a within binding conflicting scenario

| | Political orientation | | | | |
|---|---|---|---|---|---|
| | Very liberal | Liberal | Centre | Conservative | Very Conservative |
| **Conflict** | 0.11 | 0.12 | 0.17 | 0.30 | 0.31 |
| **MD*** | 0.25 | 0.31 | 0.47 | 1.06 | 1.14 |

**Moral Disengagement.* There is a downward trend from right (very conservative) to left (very liberal) for both probability of conflict and moral disengagement expected utility.

within a binding conflict event. But what about a conflicting event *between* individualising and binding foundations? That seems less straightforward, but exploiting PGM framework features make such queries simple. Running diagnostic reasoning on a scenario where we have observed a potentially conflicting event between individualising and binding foundations $P(c^1 \mid MF_e[Left^2, Left^2, Right^2, Right^2, Right^2], PO)$ we get the probability of conflict for each political orientation 0.35, 0.37, 0.38, 0.37, 0.33, and the moral disengagement expected utilities 1.5, 1.6, 1.66, 1.61, 1.34 (see also table 2.10). Here we see an interesting, and unexpected, pattern. The model predicts that agents in the centre of the political spectrum are more likely to experience conflict and to morally disengage from it. Although moderate liberals and conservatives have almost the same profile, very liberal and very conservative agents differ in that extreme liberals are predicted to be more prone to experience conflict and morally disengage. Note here that this time the results do depend on which combination of foundations we use. For example, the result would be different if we had a scenario where only one individualising foundation was enabled, against all the three binding foundations $MF_e[0, Left^2, Right^2, Right^2, Right^2]$. We report all possible meaningful combinations in figure 2.6, but we spend more time

Table 2.10
Probabilities of conflict, and expected utilities for moral disengagement for a between individualising and binding foundations scenario

| | Political orientation | | | | |
|---|---|---|---|---|---|
| | Very liberal | Liberal | Centre | Conservative | Very Conservative |
| **Conflict** | 0.35 | 0.37 | 0.38 | 0.37 | 0.33 |
| **MD*** | 1.5 | 1.6 | 1.66 | 1.61 | 1.34 |

**Moral Disengagement.* In both the probability of conflict and the moral disengagement expected utility, there is an upward trend starting in very liberals and peaking in centre individuals, but after that, the increase fades away.

discussing the *maximal* case where all of the foundations are enabled, and the two clusters point to different directions (see figure 2.5, BIBC).

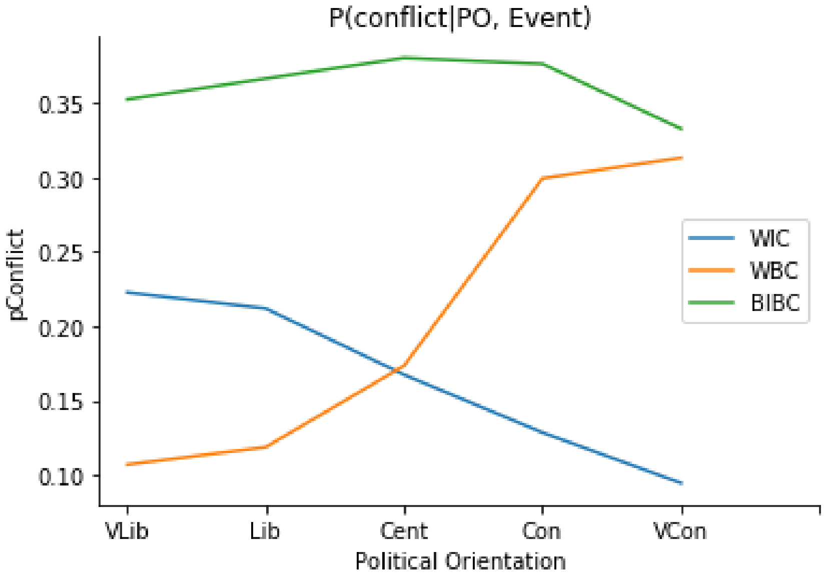


*Figure 2.5.* Moral disengagement expected utility over political orientation. *WIC*: Within Individualising Conflicting event. *WBC*: Within Binding Conflicting event. *BIBC*: Between Individualising and Binding foundations Conflicting event.

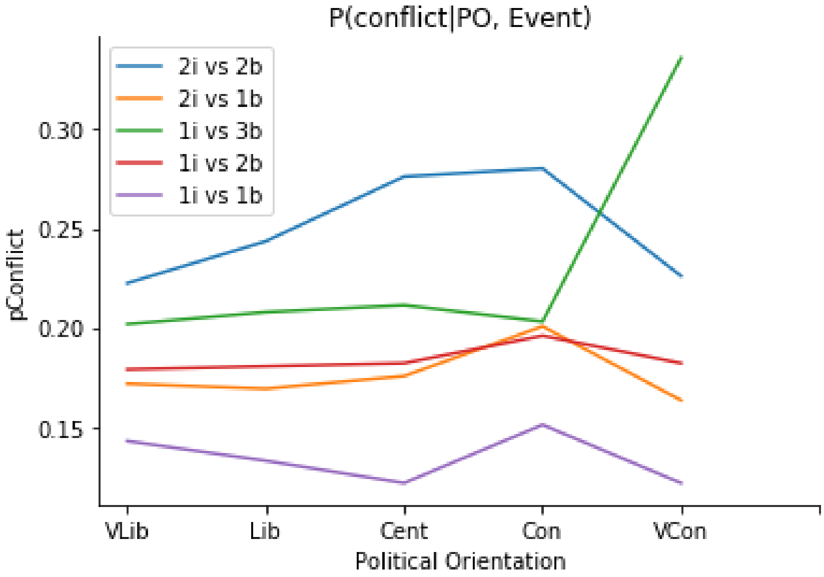


*Figure 2.6.* Probability of conflict over political orientation.

Another interesting pattern arising from the above results is that of the magnitude of conflict for each of the conflicting scenarios. We can see that BIBC triggers the greater probability of conflict, regardless of political orientation, followed by WBC and then WIC. This is also confirmed by running diagnostic reasoning on the conflict variable, but this time having only observed the event's moral foundations leaving aside political orientation: 0.17, 0.20, 0.36, for WIC, WBC, and BIBC, respectively.

## 2.4 Discussion

The aim of the current work was to explore how internal conflicts between moral values might lead to moral disengagement. We investigated how frequently agents experience moral conflict, and in turn disengagement, as a function of moral values and political orientation. For the most part, the findings aligned with our expectations: the more extreme someone is on the political spectrum, the more probable they are to experience conflict, and more keen on disengaging as a means to resolve the conflict they feel. This finding extends Critcher et al.'s (2009) results in that it provides evidence supporting that liberals and conservatives differ in multiple ways in their experience of moral conflict (*ideological inconsistencies* in their study). Specifically, they suggest that conservatives are less likely to regard their own inconsistent positions as inconsistencies. That can potentially be explained by our finding that conservatives are more prone to disengage from their conflicting beliefs than liberals are. This, in turn, could make conservatives less likely to perceive their incongruent positions. It is worth mentioning that these findings do not align with criticism of MFT supporting that MFT's definition of harm is too ambiguous. In particular, the Theory of Dyadic Morality (TDM) supports that harm, and more generally morality, is in the eye of the beholder, thing that eliminates our ability to talk about conflicting beliefs altogether. The definition of conflict used in this dissertation is that "two beliefs are dissonant if the obverse (i.e., opposite) of one follows from the other", which implies a degree of subjectivity in these beliefs. In other words, the individual cannot experience conflict if they do not perceive their two or more beliefs as conflicting.

An unexpected finding was the sudden decrease in conflict and moral disengagement in extreme conservatives, relative to conservatives. Specifically, as we mentioned above, the main trend of the results indicates that the further right one scores on the political spectrum, the more probable is conflict, and in turn, moral disengagement. However, extreme conservatives break that trend by being less likely to experience conflict than conservatives. The literature does not mention any relevant dissimilarity between extreme conservatives and conservatives, however, inspecting Graham et al.'s (2009) findings (also depicted in figure 1.1), suggests a possible explanation. Within our model, a conflict requires both strongly and equivalently upheld beliefs. A bit more formally, the difference between the two (or more) conflicting foundations works as a weighting parameter for the endorsement of the foundations. That is, the greater the difference, the less the conflict, as one (set of) moral value(s) is more influential than its competitors. Applying this to conservatives (see figure 1.1), we see that the *difference* between any pair of the foundations should not be high since all foundations are *equivalently* valued. On the other hand, extreme conservatives value the binding foundations (loyalty, authority, sanctity) more the the individualising ones. Therefore, in any BIBC, the difference between the foundations should be greater (compared to conservatives), which in turn decreases the probability of conflict.

This is an unexpected and interesting result. The model clearly predicts that extreme conservatives differ from conservatives in a way different from how extreme liberals differ from liberals. Such a result is novel, but there are related findings of psychological asymmetries as a function of ideology in the wider literature (e.g., J. Baron & Jost, 2019; Deppe et al., 2015; Jost, 2017; Pennycook & Rand, 2019; Yılmaz & Sarıbay, 2016). Specifically, there is an ongoing debate on how partisan bias is distributed across the political spectrum. On the one end there is the *symmetry hypothesis* (e.g., R. A. Altemeyer & Altemeyer, 1981; R. A. Altemeyer & Altemeyer, 1996; Eidelman et al., 2012; Wojcik et al., 2015) which predicts that levels of partisan bias will not differ between liberals and conservatives, and on the other end there is the *asymmetry hypothesis* (e.g., Alicke, 1985; Mercier & Sperber, 2011; Darley & Gross, 1983) which predicts greater partisan bias in conservatives than liberals. Our finding (i.e., the asymmetric trend between conservatives/very conservatives and liberals/very liberals) fits better the asymmetry hypothesis. That is, the dissonant state one experiences when confronting conflicting beliefs could plausibly manifest itself as favouring information that supports rather than challenges ones political affinities (as a method to avoid the unpleasant emotion cause by conflicting beliefs). Conservatives, having a greater probability to experience conflicting beliefs, they might also have a higher tendency towards partisan bias, which is what the asymmetry hypothesis predicts.

Regarding the conflict triggered within different scenarios, the findings were more anticipated. In particular, when the enabled event's foundations are only *care* and *fairness* (i.e., individualising foundations), and they point to different directions (e.g., left/right, respectively), we have set the stage for a *within individualising foundations* conflict. A contemporary example of such a conflict is ongoing debates on university campuses regarding 'hate speech' and 'free speech'. Here, there is arguably a tension between valuing the prevention of harm that hate speech might produce and ensuring the fairness of open dialogue and exchange of ideas that might otherwise be unpopular. For someone who strongly values both care and fairness, this can be a significant conflict event. The model results support this position: in a within individualising foundations conflicting scenario, the more one values these foundations, and the farther to the left one scores on the political spectrum, the more probable it is to experience conflict. This extends to disengagement as well, with liberals and extreme liberals being more willing to disengage than conservatives and extreme conservatives. When the enabled event's foundations are three binding foundations (*loyalty*, *authority*, and *sanctity*), and point to different directions, the results are again straightforward. Political agents who strongly and equivalently value these foundations are more likely to experience conflict and disengage. In other words, the further to the right one scores on the political continuum, the more likely they are to experience conflict. One thing to note here is that the fading-off pattern of extreme conservatives, observed in the main results where the event's moral foundations are left to vary, is absent here. In other words, extreme conservatives are *more* likely to experience conflict than conservatives in a within binding foundations conflicting scenario. We see that when the individualising foundations are turned off, the probability of conservatives to experience conflict decreases more than that of extreme conservatives. That is, *care* and *fairness* foundations can explain at least some of the conflict probability in conservatives, while almost none in extreme conservatives. That is to be expected, since conservatives value more equally the five moral foundations than extreme conservatives.

We saw that liberals are more likely to experience conflict in a within individualising foundations conflicting event, while conservatives are more probable to experience conflict in a within binding foundations conflicting scenario. One last point of discussion is the difference in probability between these two different settings. Specifically, extreme conservatives are more likely to experience conflict in a within binding conflicting event than liberals are in a within individualising conflicting event. This finding, besides attesting to the higher probability of conflict conservatives have, also indicates that this pattern holds true even when we control for within individualising foundations. In other words, conservatives are more likely to experience conflict than liberals, even when the within binding foundations are deactivated. This time, we can explain this solely by looking at the number of moral foundations strongly valued by each of the political ideologies. For liberals, this holds true only for care and fairness, while for conservatives for loyalty, authority, and sanctity. Therefore, in this setting conservatives are more likely to experience conflict than liberals because they have more sources of beliefs than liberals. After making a few simplifying assumptions, such as that each of the moral foundations contains the same number of beliefs, our result becomes as straightforward as saying that three is more than two. Though we haven't explored relaxing this assumption here, it is worth exploring briefly. As discussed previously (internal section citations), there is strong evidence for the link between political ideology and moral values across a wide range of cultures. Similarly, political ideology is generally evenly distributed across the population (i.e. there are approximately equal numbers of left and right wing adherents; citation needed). If we allow different moral foundations to contain differing numbers of beliefs, this directly implies that events in the world can have differing probabilities of triggering/being relevant to those beliefs. The question arises: how many more beliefs would individualising foundations need to have, compared to binding foundations, to render the probability of an event triggering either set equal[8]? The answer is approximately 1.5. This attempt to render individualising and binding events equi-probable necessitates making individualising foundations substantially more relevant to the world at large, which is an extremely contentious move (but see Gray & Keeney, 2015a, 2015b; Gray et al., 2014; Gray, Waytz et al., 2012; Schein & Gray, 2018). In other words, we haven't set up our model to 'prove' that conservatives experience more conflict - that is simply a straightforward consequence of assuming that all moral foundations are equally relevant to the world, regardless of their association with any particular political ideology.

Things become a bit less straightforward when we set up a *between* individualising and binding foundations conflicting scenario. In the most extreme such scenario, all five moral foundations are enabled, with individualising pointing to one direction and binding to the opposite. The results of this case are interesting. Conflict probability peaks over the centre of the political spectrum and then falls off toward both extremes at roughly the same rate. But why is the probability of conflict in this setting wildly different than in the case where none of the event's moral foundations have been observed? The answer actually reveals a weakness of our model, and more generally of Bayesian modelling: the influence prior values have on the outcome. To remind you, we used a uniform distribution over the event's moral foundations being enabled at various levels, while we gave a lower probability for the case where an event's moral foundation is disabled. Having used different priors, perhaps ones where the probability of being enabled is higher for some foundations, the results would have looked more like the ones of a BIBC scenario. However, our choice of priors is justified on the grounds that we do not actually posses the information of how frequent an event's moral nature is salient and in what way, so for that reason we use a uniform distribution over all enabled levels. Moreover, the difference we observe between the case where all foundations are observed to be enabled and each cluster points to different sides (figure 2.3, and the case where none foundation is observed to be enabled (thus they have an equal probability of being enabled or disable –figure 2.5 green line), can be explained by the fact that, in the former case, we exclude

[8] Equivalent to asking how many fewer beliefs would binding foundations need to have.

the possibility of a conflict within each cluster, which in turn decreases the probability of conflict among the two extremes (since we saw that liberal are more likely to experience conflict in a *within individualising foundations* conflicting event, while conservatives in a *within binding foundations* conflicting event.

### 2.4.1 Theoretical implications

This research brings together the moral foundations theory (Haidt & Joseph, 2004, 2007; Graham et al., 2009; Haidt, 2012), dissonance theory (Festinger & Carlsmith, 1959; Festinger, 1962), and moral disengagement theory (Bandura et al., 1996; Bandura, 1999), into a theoretically explicit probabilistic graphical model (Koller & Friedman, 2009). Our results present what we believe is the first formal method for predicting conflict and disengagement frequency, given the political orientation of the agent, the agent's moral values, and a given social setting/morally relevant event. To our knowledge, this model is first of its kind, although it joins with prior approaches to modelling political belief systems (Brandt et al., 2019) and individual attitudes (Dalege et al., 2016) to show the value of computational modelling in the social sciences. In doing so, we have provided additional detail about the ways that specific event settings trigger different types of conflict.

Building on earlier work, our findings suggest substantive differences not only in liberals' and conservatives' average endorsement of moral foundations items (Graham et al., 2009), but also in liberals' and conservatives' probability of experiencing conflicting beliefs. It adds to the current research by presenting an unprecedented finding of differences between the two edges of the political spectrum. In particular, we could not find any research reporting that the relationship between liberals and extreme liberals is different than that between conservatives and extreme conservatives. The common case in the literature is to compare liberals to conservatives like it was done in Turner-Zwinkels et al. (2020), where they compared these two political ideologies and found that they have different systems of foundational moral values, with liberals' moral systems showing more segregation between individualising and binding foundations than conservatives.

So far we have discussed conflicts *within* a political agent, but, in theory, our model is not restricted to explaining only this type of conflict. Specifically, it can make predictions relevant to just the event's moral foundations, that is, it can make predictions on the probability of conflict –be it within agent or between agents–, in a particular moral setting. The model predicted that the most fertile ground for a conflict is in a between individualising and binding foundations conflicting scenario. That finding could have implications for *echo chambers* and *polarisation* research, where people seem to surround themselves with others who have the same or compatible beliefs, in an attempt to avoid conflict, and reflect and reinforce their own beliefs (e.g., Garrett, 2009; Farrell, 2015). As far as *conflict avoidance* goes, extending our findings to this line of research we would expect to see people creating "bubbles" based on the two clusters of foundations (individualising and binding), and that is because the biggest chance a scenario has to trigger a conflict, is when foundations between the two clusters are enabled and pointing towards different directions. In other words, in order to avoid conflict with other people, individuals who uphold individualising foundations may withdraw themselves from exposure to those who uphold binding foundations (and vice versa; cf. (see for example Baumann et al., 2020): liberals and conservatives tend to create their own echo chambers and "live in their own bubbles").

In an attempt to avoid conflict with other people individuals create their own environment with like-minded others. Taking that a step further by looking at our findings, we would expect liberals to form more dense echo chambers, meaning that they would be less probable to drift to others' ideological viewpoint, than conservatives. That is because, according to our model predictions, conflict is *less* probable in a within individualising foundations scenario than it is in a within binding conflicting scenario. In other words, conservatives are more probable to clash with one another than liberals, leading to less dense echo chambers. That theorising is partly supported by Eady et al., 2019. Eight five percent of liberals choose not to be exposed to an ideological opinion (Fox News) that more than 66% of extreme conservatives choose to. By contrast, 78% of conservatives choose not to be exposed to an ideological view (MSNBC) that 78% of liberals choose to. Therefore, whereas there is significant overlap in what both groups view, there are also zones of the media eco-system that are mainly seen by individuals of one ideological group, with conservatives being slightly more probable to expose themselves to different ideological viewpoint than liberals. In chapter 5 we use an agent-based model to investigate further whether this prediction holds true, and what might be other factors contributing to splitting people apart.

### 2.4.2 Known limitations

The present study offers a cross-sectional investigation of moral systems in different political parties. Although it gives a lot of potentially useful insights on how frequently people experience conflict and choose to disengage from their beliefs, it also carries some methodological disadvantages. A theoretical point regarding the structure of the model concerns the loose use of the notion of causation. Specifically, the general approach we adopted is to choose a structure that reflects the causal order and dependencies, so that causes are parents of the effect. However, that does not always hold true in the model, with the only case being the relationship between political orientation and agent's moral foundations. Although the political orientation node is parent to the five moral foundations, implying in that way a causal relationship between them, in actuality we do not take position in that debate. That is, we do not claim that political orientation *causes* an agent's moral values, nor the opposite. In model's defence, as we mentioned section 2.2.3, the causality is in the world, not in the inference process. That is, there is no statistical implication following from that, only a theoretical one.

Another way we could extend our model would be to incorporate an optimisation function making the model capable of fitting to observed sample characteristics. However, that does not mean that our model findings are not valid or inaccurate, just that the process of acquiring the results is more cumbersome and time consuming to carry out.

A simplifying assumption we made in this version of the model is the fact that we do not differentiate between foundations belonging in the same cluster (individualising and binding). We regard both individualising foundations as being the same, giving them the same weights (CPDs), as well as all three binding foundations have the same weights (see also appendix A.3). We could have given each foundation its own weights, adding to the level of detail the model has, but we chose simplicity over granularity. However, we do amend for that in the next version of our model, presented in chapter 3), where we introduce differing CPDs for foundations within clusters.

Another theoretical modification we could apply to make the model a better surrogate of moral foundations theory is to add potential interactions between moral foundations. There are substantive correlations between foundation endorsements (Haidt, 2012). For example, an agent who upholds the care foundation is more likely to also uphold the fairness foundation. In the current version of the model we have not incorporated that property yet, but in the next study (chapter 3), we do (not in the foundation level but in the cluster lever, that is, binding foundations should be more likely to be enabled/disabled together, as well as individualising ones).

There are a few other limitations which could lead to potential upgrades regarding the technicalities of the model. For example, a potential upgrade could be the granularity level of the moral disengagement node. At its current state, moral disengagement node has only two levels: either you disengage or not. First, a possible extension would be to introduce from which of the beliefs the agent should disengage in order to maximise the expected utility, and second, we could increase the level of detail with regards to moral disengagement itself. In the literature exist at least eight mechanisms some of which being intertwined with one another, while in the model we regard moral disengagement as a binary variable.

To conclude, the current study demonstrates probabilistic graphical models are valuable when modelling systems having inherit uncertainty, like the frequency of conflict within political agents. In doing so, we bring together influential social and cognitive psychology frameworks to explore the frequency of conflicting moral beliefs and moral disengagement in different political ideologies. We have presented novel evidence that conflict frequency is different between different political ideologies with left-ish positions being less probable to experience conflict and disengagement. Though the current version of the model can be extended in various ways, we still believe it lays the foundations for a potentially fertile line of research which has interesting implications.

# Chapter 3

# Modelling the frequency of moral disengagement on the political continuum: version 2

## 3.1 Introduction

Most common in Moral Foundations Theory (MFT) literature are studies investigating differences across the political spectrum in values and moral judgement. These differences, as we also saw in chapter 2, can explain some of the variation of how frequently political agents experience conflict and, as a result, (temporarily) disengage from their beliefs. In particular, assuming that the endorsement of a moral foundation does not directly affect the endorsement of another, and that moral foundations within moral foundation clusters (individualising/binding) are equally probable to be either conflicting or consonant, we found that the more on the left one individual is on the political spectrum, the less probable they are to experience conflict and disengage from their beliefs. However, the endorsement of one foundation does empirically correlate with the endorsement of another (Haidt, 2012; Simpson, n.d.), for example, if one endorses sanctity foundation then one is more likely to also endorse loyalty than care foundation. Furthermore, we can theorise that, since foundations within foundation cluster are normally upheld together, the beliefs belonging to these foundations may also be more likely to be consonant with one another. That is, beliefs coming from within a cluster of foundations could be less likely to conflict on average, leading to less within cluster conflicts than between cluster conflicts. To explore this possibility, we relax the assumption of interaction between foundations, letting foundation endorsements co-vary in a direct manner (i.e., the endorsement of one foundation is directly linked to the endorsement of another), and we restrict the assumption of agreement within clusters of foundations forcing a higher probability for non-conflicting beliefs within the same cluster. Building on the model presented in chapter 2, we believe that these small but significant changes will increase model accuracy and introduce little to non-existent complexity.

### 3.1.1 Interaction within clusters of foundations

Moral foundations theory arose within social psychology originally as a way to extend Shweder et al.'s (1997) research and to describe the psychological components in ethical disputes, political beliefs, and cross cultural variance in moral judgements. The individualising moral foundations (care/harm and fairness/cheating[1]) derive from *ethic of autonomy* of Shweder et al. (1997). They are mainly tuned to persons as subjects of (instead of agents of) ethical activity, and are enabled over impression of people's plights, sufferings, needs, rights, and welfare. In other words, these foundations treat individual persons as the fundamental moral unit of consideration. Care is propelled towards matters of compassion, empathy, sympathy, nurturance, and harm decrease, and is fired when one detects violence, aggression, emotional abuse, bullying, or, more broadly, suffering. Fairness is geared towards matters of justice, fairness, proportionality, equality, and reciprocity. It is enabled when one observes injustice, disproportionate distributions of goods, merits, favours, benefits, punishments etc., inequality deemed excessive, and failure to reciprocate. Either foundation is favoured in a greater degree by left-wing, compare to right-wing, proponents. Furthermore, there seems to be a positive co-variance between the endorsement of the two foundations (Graham et al., 2009; Haidt, 2012).

[1] Hereafter only the virtue of each foundation name will me mentioned.

The binding foundations (*loyalty/betrayal*, *authority/subversion*, and *sanctity/degradation*) serve to reinforce social and interpersonal coherence and to defend the community. In that respect, they are mainly tuned to people as agents of (instead of subjects of) ethical activity, who are able of both backing and weakening the ingroup regarding its integrity, inter-group standing, tradition, order, and sanctity. Here, the fundamental moral unit of concern seems to be the group/tribe, rather than the individual. Loyalty and authority foundations derive from Shweder et al.'s (1997) *ethic of community.* Loyalty promotes matters of union, solidarity, collective, culture and shared identities, families and tribes. Possibly emerged in the interest of inter-group competition (e.g., supporting/encouraging towards members of the same group), this foundation entails an ethical duty to defend and favour the ingroup, and is activated in the face of disloyalty, betrayal, excessive individualism, or threats to ingroup identity and coherence (e.g., as some may regard immigration). Authority gears towards matters of order, hierarchies, authorities, traditions, and respect for persons of higher power/status, and is triggered in the face of subversion, disrespect, nonconformity, or disobedience. Last, equivalent to Shweder et al.'s (1997) *divinity*, sanctity orients towards defending the dignity, purity, cleanliness and sacredness of the ingroup. Acts deemed in breach of the ingroup's nature and religious character (actions normally considered disgusting) activate this foundation, as it acts to defend the ingroup from contamination, incorporating perceptions of physical (e.g., from illness), spiritual (e.g., from curses of witchcraft), and social (e.g., from immigration, tribal intermarriage, premarital sex) contamination. The binding foundations exhibit a positive relationship amongst themselves, similar to the individualising foundations (Graham et al., 2009; Haidt, 2012).

Besides the correlations between individual foundations, correlations do exist between clusters of foundations. Although we do not explicitly incorporate these correlations in our model –we believe they will emerge on their own– we briefly discuss them here. In general, there is a negative relationship between the two clusters such that when one cluster is activated (i.e., level of endorsement is high), the other's activation is low. This relationship, however, holds true more for liberals than for conservatives who endorse both clusters almost equally.

### 3.1.2 Agreement between clusters of foundations

From the previous summary, we can theorise that the *agreement* between beliefs within a cluster would be higher than those from different clusters. We use the term agreement to describe the relationship between beliefs and/or foundations. For example, in our model, two beliefs that agree are two beliefs which are both activated, and both pointing towards the same direction. Disagreement reflects the opposite: two beliefs that are enabled but pointing towards opposing directions. The agreement of within cluster beliefs should be higher as foundations within cluster are involved with similar functionalities. In particular, as we saw, the individualising foundations are activated during perception of individuals' plight, suffering, needs, rights, and welfare, while binding foundations are activated by perceptions of supporting (or violating/threatening) the ingroup's integrity, intergroup standing, traditions, order, and sanctity. Thus, in the current version of the model (from now referred to as PDM-A for Probabilistic Disengagement Model Agreement) we also take into account the agreement of foundations in conflicting scenarios.

### 3.1.3 Current study

While the probabilistic graphical model (PGM) introduced in chapter 2 provided valuable insights regarding conflict within and between political agents, it relies on a few assumptions which are not supported by literature. The contribution of the current work includes two distinct model updates that relax some of these assumptions. The first revision lets the moral foundations of the agent to co-vary such that the endorsement of one foundation can influence that of another. The second revision introduces the concept of within-cluster agreement. This also involves co-variation between foundations but this time between the foundations of the event - that is, we assume that events likely to trigger a foundation within a particular cluster are likely to trigger multiple such foundations. To allow for these changes we need to introduce a slightly different type of PGM, one that does not ascribe directionality to the interaction between some variables. Hence, we implement a *partially directed acyclic graphs* (PDAG) which is capable of hosting both types of edges. Technical details about undirected graphical models and PDAG can be found in appendix A.2.2.

Let us now consider the implication of these changes. For one, we do not expect to see any significant differences regarding the pattern of results. That is, we still expect to find right-wing ideologies more likely to have conflicting beliefs than left-wing ones. Where we do expect to find differences is at the specific scenarios and the magnitude of the conflict/moral disengagement. In short, we expect to find the same pattern of results, but not the same numbers. In particular, for liberals we anticipate a lowered conflict probability when in a within individualising foundations conflicting scenario, while for conservatives a lowered conflict probability for within a binding foundations conflicting scenario. We expect to see an increased marginal (i.e., across all scenarios) probability of conflict for political centrists. We also expect to see an increased marginal (i.e., across all political orientations) probability of conflict for between individualising and binding foundations conflict scenarios. Similar patterns should arise regarding the moral disengagement variable.

## 3.2 Methodology

In this section we will only discuss the revisions and not the whole model. The interested reader can find a detailed description of the complete model in section 2.2. Moreover, the interested reader might also want to refresh his/her knowledge on PGMs (appendix A.2 or –for a more extensive read-through– Koller and Friedman (2009)).

### 3.2.1 Theoretical description

The revisions of the PDM-A took place in the *political orientation* space which consists of the political orientation variable and the five moral foundations of the agent, and in the *conflict* space. The five moral foundations now have pairwise connections with their neighbouring nodes. That is, *care* is now connected to *fairness*, *fairness* to *loyalty*,

*loyalty* to *authority*, and *authority* to *purity*. Note that in reality each foundation is interconnected with all of the other foundations resulting in $5 \times 5 - 5 \times 1$ connections, but adding all these links in the model would increase the complexity of the model even more perhaps leading to intractable distributions. In theory, all these links just reflect that the activation of one node can influence the activation of another.

To make sure we will not unnecessarily increase the already increased model complexity, we registered the second revision in a slightly different way. This time we did not add connections between the event's moral foundations, instead, we redesigned the conflict function. As we will see, this has technical and theoretical implications. We introduced a mechanism which intensifies conflict based on whether the event's moral foundations belong in the same cluster or not. For example, if, in a given event, there are two foundations enabled and conflicting, and each of which is coming from a different cluster (i.e., individualising vs binding) then we boost the probability of conflict by a small increment. Adding this mechanism in the conflict function implies that the event's moral foundations do not explicitly influence one another, rather, the influence takes place within the agent. We theorise that this is the case as the nature of the event's moral foundations is rather subjective, and that the agent is the one who ascribes meaning to them. This is also the case for the agent's moral foundations: the influence of one on another takes place within the agent. After the calculation of the agreement parameter, and the subsequent revision of the conflict, the computations of the moral disengagement remained unchanged. You can see the PDM-A graph in figure 3.1.

### 3.2.2 Formal description

The model has now a few important modifications regarding its formal qualities. First of, the addition of undirected edges mutates the model from Bayesian Network to Partially Directed Acyclic graph (PDAG). As we discuss in appendix A.2.3, PDAGs have a few different properties from directed graphs. Perhaps the most important difference is the introduction of pairwise factors which mutually influence one another. That is, the activation of one is related to the activation of the other. Note that there is not a parent-child

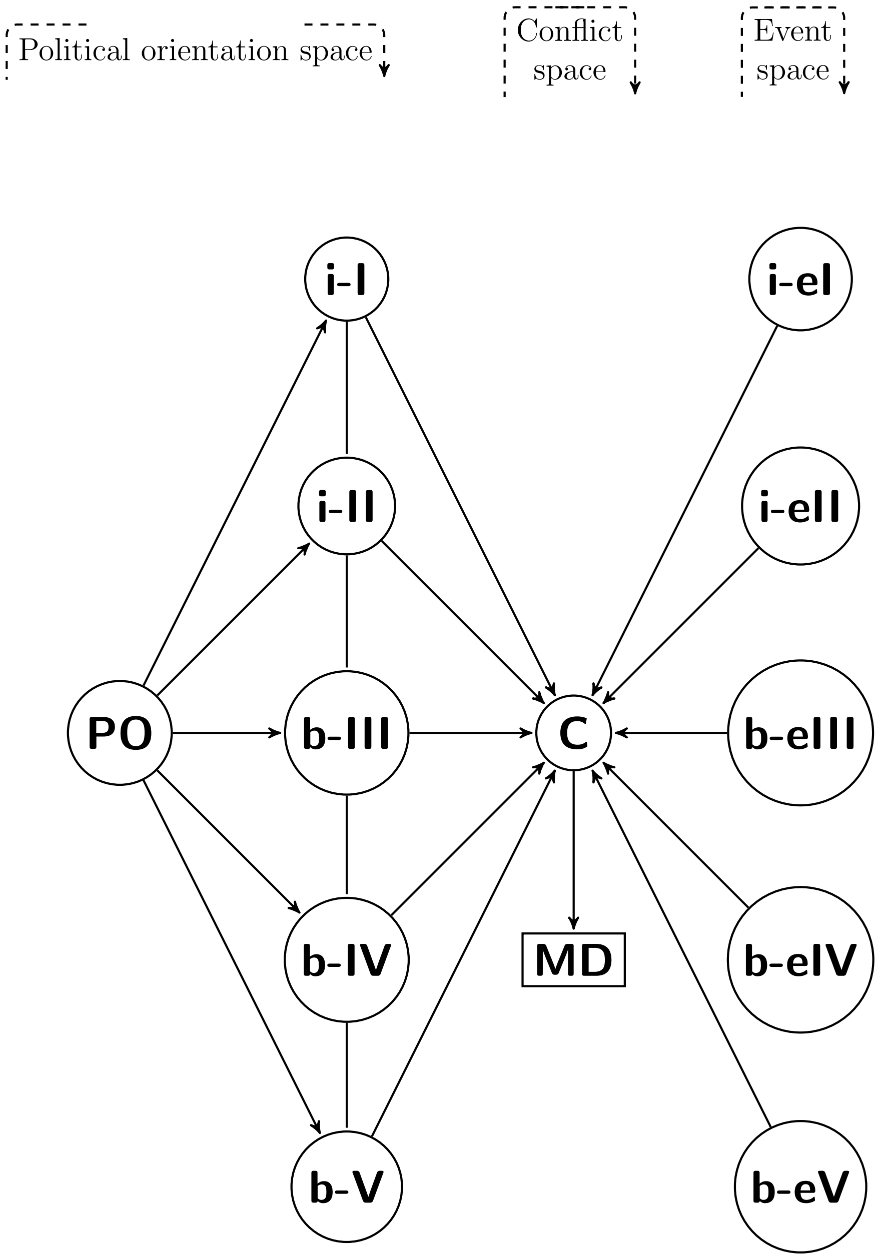


*Figure 3.1.* The revised version of the model having the undirected edges between the agent's moral foundations added. $\mathcal{G}$

relationship between them, rather, they are simply two neighbouring nodes connected by an undirected edge. Formally, the edges we added are: $mf_a^1$—$mf_a^2$, $mf_a^2$—$mf_a^3$, $mf_a^3$—$mf_a^4$, $mf_a^4$—$mf_a^5$. Note that we only add pairwise edges, meaning that, for example, $mf_a^1$ links only to $mf_a^2$ and not to $mf_a^3$. However, node $mf_a^1$ can still –indirectly– affect node $mf_a^3$, since node $mf_a^1$ affects $mf_a^2$, $mf_a^1$—$mf_a^2$, which, in turn, affects $mf_a^3$, $mf_a^2$—$mf_a^3$. Conveniently, as far as *agent's* moral foundations go, the conflict function remains the same since the levels of each of the agent's moral foundation variables are the same (i.e., disabled, enabled low, enabled medium, enabled high). The conflict function does change, however, when it comes to event's moral foundations.

Instead of using undirected edges and factors to portray relationships between an event's moral foundations, we do so in the conflict function. There are both theoretical and technical motivations for this. Technically, adding more factors would render the model inefficient if we wanted to add learning to it, as the parameterisation would explode, and learning across such a large parameter space would require more data than is available to us. Moreover, from a theoretical perspective, we prefer locating the relationships within the agent (i.e., in the conflict function), rather than the external world (i.e., event's moral foundations), as we believe the agent is the one who ascribes moral value to the event; thus, psychologically, the influence process should be intrinsic. Formally, the conflict function is now:

$$P(C \mid MF_a, MF_e) = 1 - sqrt(e^2) * d + exp(a; \beta) \tag{3.1}$$

where $e$ is the *energy*, or how enabled all foundations are, $d$ is the total difference between the foundations, $a$ is the *agreement* metric between the foundations, and the $exp(\chi; \beta)$ is an exponential function on $\chi$ with a *scale* parameter $\beta$. We calculate $e$ like so:

$$e = \frac{\hat{f}}{\max f * 0.5} \tag{3.2}$$

where $\hat{f}$ is the mean of the left, $l$, and right[2], $r$, sums of the agent's foundations multiplied

[2] Note again that left and right are not political terms, but rather indicate disagreement between active foundations.

by the event's foundations resulting in pointing either to the *left*, or the *right*:

$$\hat{f} = \frac{l + r}{2} \tag{3.3}$$

We scale $\hat{f}$ with the constant $\max f$ which is the maximum value possible out of the sum of $l$ and $r$. This constant of course depends on the given parameters; in our case $\hat{f} = 15$. We halve this value as the maximum *mean* of the *left* and *right* sums cannot exceed $\max f/2$.

The difference $d$ in equation 3.1 is computed by calculating the difference between the left and the right activation values and dividing this by $\max f$ like so:

$$d = \frac{l - r}{\max f} \tag{3.4}$$

We do not have to halve $\max f$ here, since, given the initial parameters, the value of the difference can exceed $\max f/2$ but not $\max f$.

The agreement term $a$ is computed by calculating the distance $mu_D$ between the mean activation of the individualising foundations $mu_i$, and the mean activation of the binding foundations $mu_b$, and scaling by 6 which is the max distance between $mu_i$ and $mu_b$. Then we multiply this distance by the complement of the distance each of the clusters has from zero $z_D$. The bigger the distance between the moral foundation activations, the less the agreement. After having calculated the agreement, we plug it into an exponential distribution function and add the results to the conflict. Larger agreement values result in smaller increments of the exponential function. Formally, the agreement parameter is computed like so:

$$a = \frac{mu_D * (1 - z_D)}{dP} \tag{3.5}$$

where $dP$ is a constant parameter which controls the magnitude of the increment.

### 3.2.3 Initial setup

In general lines, the model set up was the same as its first version in chapter 2. The principal change is the incorporation of connections between the agent's moral foundations, in the form of pairwise factors. These connections are fixed throughout model iterations. Since our model is not equipped with the capacity to learn, we manually added correlation weights between foundations based on $689,384$ empirically observed values (Moral Foundations Data Set from YourMorals.org, personal communication Haidt, 2018). YourMorals.org website, which has over the years collected self-reported responses from over 200,000 respondents and contains subsamples of thousands or tens of thousands of respondents from different global regions Briefly, we added positive correlations within foundation clusters (i.e., individualising and binding), and a negative correlation between clusters. Since we only used pairwise correlations, the specific links were: positive for $mf_a^1$—$mf_a^2$, $mf_a^3$—$mf_a^4$, $mf_a^4$—$mf_a^5$, and negative for $mf_a^2$—$mf_a^3$. Besides the incorporation of these correlation weights the rest of the model setup remained the same.

## 3.3 Results

### 3.3.1 Probability of conflict for political agents

The main question the current study remains the same: whether there is any difference in the conflict and moral disengagement frequencies between different political orientations. The way we tackle this is again by using diagnostic reasoning having first observed some evidence. Comparing just the probability of a conflict between conservatives and liberals we have $P(c^1 \mid liberal) \neq P(c^1 \mid conservative)$, and in particular $P(c^1 \mid liberal) < P(c^1 \mid conservative)$. As we can see from table 3.1, when we hold everything else equal, liberals' probability of experiencing a conflict is 0.28 while conservatives' probability of conflict is 0.34. It is important to note here that none of the event's moral foundations have been observed - they follow a uniform distribution for the four levels of being activated (activated low and pointing to the left, activated high and pointing to the left, activated low and pointing to the right, activated high and pointing to the right). While there is a small chance of not being activated at all, when we have a political agent in potentially conflicting settings, then if this political agent is liberal, she is less likely to experience conflict than if she is conservative.

Moving away from the centre, further to the right and to the left of the spectrum we

Table 3.1
Probability distributions over $P(C \mid PO)$

| | | **Political orientation** | | | |
|---|---|---|---|---|---|
| | | V.Liberal | Liberal | Conservative | V.Conservative |
| **Conflict** | Disabled | 0.74 | 0.72 | 0.66 | 0.66 |
| | Enabled | 0.26 | 0.28 | 0.34 | 0.34 |

*Note.* We can see that liberals are less prone to conflict than conservatives, keeping everything else equal.

have extreme liberals and extreme conservatives. Repeating the same steps and applying the same algorithms for this comparison, we find that extreme liberals' probability of experiencing a conflict is 0.26 while extreme conservatives' probability of conflict is 0.34 (see table 3.1). These results are quite interesting. First we can see that the model predicts that conservatives have a higher marginal probability of conflict than liberals, but what is more interesting is the relative relationship of this trend to the trend depicted in the conservatives versus liberals table 3.1. Comparing conservatives to extreme conservatives we observe that these two ideologies are almost equally probable to experience conflict, while extreme liberals are slightly less probable to experience conflict than liberals. The upward trend of conflict from left to right breaks at conservatives, where extreme conservatives display a similar chance of experiencing conflict. Table 3.2 depicts the differences between these two groups of political ideologies (extreme and less extreme).

Table 3.2
Difference between probability distributions over $P(C \mid \mathit{extreme}^*)$ and $P(C \mid \mathit{standard}^{**})$

| | | **Political orientation** | |
|---|---|---|---|
| | | Very liberal - Liberal | Very conservative - Conservative |
| **Conflict** | Disabled | 0.2 | 0.00 |
| | Enabled | -0.2 | 0.00 |

*Note.* **Extreme* represent the two extreme political ideologies: very liberals and very conservatives; *Standard*** represent the two less extreme political ideologies: liberals and conservatives. We can see that very liberals probability of conflict is less than this of liberals', but the right-wing ideologies are equally probable to experience conflict.

Let us now turn our attention to moral disengagement. As a reminder, a decision making unit merely reflects the *willingness* of an agent to take an action, that is, the relative order and magnitude of the willingness to morally disengage. Calculating the expected utility of moral disengagement actions between liberals and conservatives for the aforementioned scenario, we find that liberals' expected utility for disengaging after a potential conflict is 1.15, while conservatives' score is 1.57. That is, in general, given the potential for conflict, conservatives would choose to disengage more often than liberals.

Computing the expected utility of moral disengagement for extreme liberals and extreme conservatives reveals that the expected utility for disengaging after a potential conflict is 1.07 for extreme liberals, while extreme conservatives' score is 1.54. We can see that there is a similar trend between extreme and less extreme ideologies with regards to conflict. That is, extreme liberals are less prone to disengagement and less likely to experience conflict than liberals. Similarly, extreme conservatives less more prone to disengagement, as well as to experience conflict than conservatives (see figures 3.2 and 3.3).

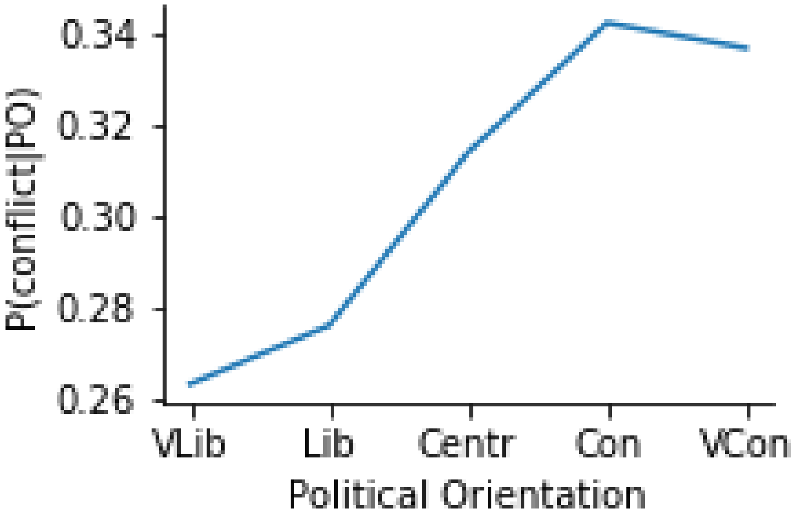


*Figure 3.2.* Probability of conflict over political orientation.

### Different moral scenarios

The main questions of our project were answered by the aforementioned queries. These queries did not involve the event's moral foundations, in the sense that the event's moral foundations were not observed. That is, these queries were focused on average agent tendencies to experience conflict and disengage, rather than their probability to do so in response to a specific event type/profile. However, we will now turn our attention to queries involving the event's moral foundations. Such queries can answer questions of

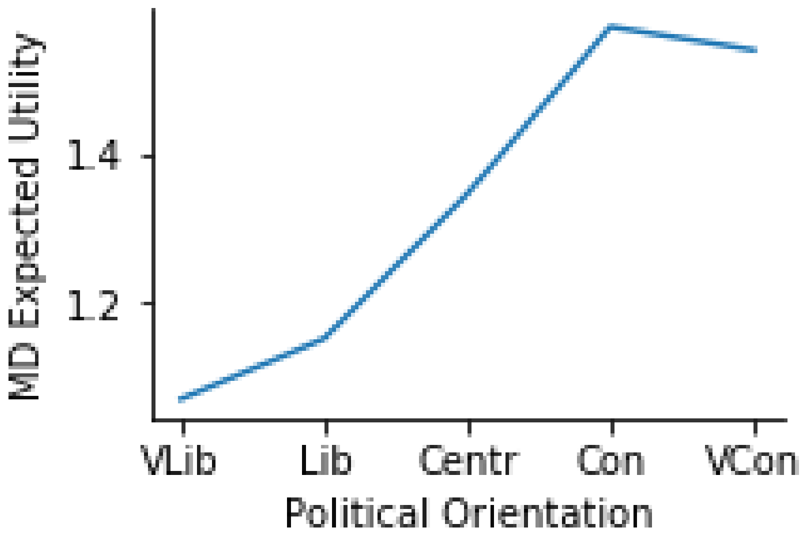


*Figure 3.3.* Expected utility of moral disengagement under different conflicting scenarios over political orientation.

the sort ”given that the political orientation of an agent has been observed, what is the probability of conflict in response to *a within binding conflicting event*”? To formalise this, we would just have to observe, e.g., a liberal, and then a conflicting scenario which involves only binding foundations. To do so we could set one of the binding foundations to be enabled and pointing to the right, and another binding foundation to be enabled and pointing to the left. Let us see this in practice.

As in the previous chapter, there are three different types of conflicting scenarios: within individualising conflict (WIC), within binding conflict (WBC), and between individualising and binding conflict (BIBC). The WIC setting is a scenario where the two individualising foundations are both enabled but pointing to different directions, while the three binding foundations are all disabled. Formally, our problem now is to find the probability of conflict given a within individualising conflicting event, for each political orientation:

$$P(c^1 \mid I_e = Left^2, II_e = Right^2, III_e = 0, IV_e = 0, V_e = 0, PO) \tag{3.6}$$

where $Left^2$ and $Right^2$ indicate that the given foundation is enabled and pointing either to the *left* or to the *right*. $Left^1$ would denote an lowly activated foundation pointing to the left while $Left^2$ would denote a highly activated foundation pointing to the left.

For the sake of simplicity, we will not mention the name of the foundation if all five foundations are observed. For instance, equation 3.6 is equivalent to:

$$P(c^1 \mid MF_e[Left^2, Right^2, 0, 0, 0], PO)$$

where $MF_e[Left^2, Right^2, 0, 0, 0]$ are the values assigned to the event's five moral foundations going from left ($I$) to right ($V$).

In terms of computations, the steps we follow to solve this query are the same as before, with the only difference being that now we have observed more evidence (we know the event's moral foundations values). Running the model with these settings, we get the probability of conflict under a WIC scenario for each political orientation: 0.39, 0.33, 0.21, 0.14, 0.09 for extreme liberals, liberals, centre, conservatives and extreme conservatives, respectively (see also table 3.3, row 1). We can see that there is a downward trend spanning from very liberal to very conservative, with very liberal having the higher probability to experience conflict while very conservative the least. That is of course to be expected, as liberals are the political agents who most uphold the individualising foundations (Graham et al., 2009), and since the conflict function requires strongly held beliefs, liberals would be more likely to experience conflict. Regarding the moral disengagement, we see that the same pattern arises, with expected utility values decreasing from left to right across the political spectrum: 1.68, 1.36, 0.82, 0.52, 0.32 for very liberal, liberal, centre, conservative and very conservative, respectively (see also table 3.3, row 2) in a scenario where only the individualising foundations are enabled.

Naturally, the next query is for the probability of conflict in a WBC scenario $P(c^1 \mid MF_e[0, 0, Right^2, Left^2, Left^2], PO)$. Note here that there are more options for conflicting scenarios since the binding foundations are one more than the individualising ones. For example, another WBC scenario could be $MF_e[0, 0, Left^2, Left^2, Right^2]$ or even with one of the three foundations being disabled $MF_e[0, 0, Right^2, Left^2, 0]$. The results are more or less the same for all combinations, thus we arbitrarily report just one of them (all three binding foundations are enabled, but one of them points to the opposite direction than the rest two). The results are quite straightforward: a downward trend spanning from very conservative to very liberal. Running the model with these settings,

Table 3.3
Probabilities of conflict, and expected utilities for moral disengagement for all three types of events

| | Political orientation | | | | |
|---|---|---|---|---|---|
| | Very liberal | Liberal | Centre | Conservative | Very Conservative |
| | **Within individualising foundations conflicting scenario** | | | | |
| **Conflict** | 0.39 | 0.33 | 0.21 | 0.14 | 0.09 |
| **MD*** | 1.68 | 1.36 | 0.82 | 0.52 | 0.32 |
| | **Within binding foundations conflicting scenario** | | | | |
| **Conflict** | 0.06 | 0.12 | 0.29 | 0.40 | 0.45 |
| **MD*** | 0.23 | 0.51 | 1.26 | 1.91 | 2.25 |
| | **Between individualising and binding foundations scenario** | | | | |
| **Conflict** | 0.36 | 0.39 | 0.45 | 0.40 | 0.32 |
| **MD*** | 1.86 | 2.15 | 2.53 | 2.26 | 1.66 |

**Moral Disengagement.*

we get the probability of conflict under a WBC scenario for each political orientation: 0.06, 0.12, 0.29, 0.40, 0.45, for very liberals, liberals, centre, conservatives and very conservatives, respectively (see also table 3.3, row 3) in a scenario where only the binding foundations are enabled. As before, the moral disengagement expected utility follows the same pattern as the probability of conflict: 0.23, 0.51, 1.26, 1.91, 2.25 (see also table 3.3, row 4).

So far, We have seen that liberals are more prone to experience conflict within an individualising conflicting event, and that conservatives are more prone to experience conflict within a binding conflict event. But what about a conflicting event *between* individualising and binding foundations? That seems less straightforward, but exploiting PGM framework features make such queries simple. Running diagnostic reasoning on a scenario where we have observed a potentially conflicting event between individualising and binding foundations $P(c^1 \mid MF_e[Left^2, Left^2, Right^2, Right^2, Right^2], PO)$ we get the probability of conflict for each political orientation 0.36, 0.39, 0.45, 0.40, 0.32, and the moral disengagement expected utilities 1.86, 2.15, 2.53, 2.26, 1.66 (see also table 3.3, rows 4 and 5).

Here we see an interesting pattern. It seems like a pyramid-shaped pattern where agents in the centre of the political spectrum are more likely to experience conflict and to disengage. Then, although liberals and conservatives have almost the same values, extreme liberals are more likely to experience conflict and disengage than extreme conservatives. Note here that this time the results do depend on which combination of foundations we use. For example, the result would be different if we had a scenario where only one individualising foundations was enabled, against all the three binding foundations $MF_e[0, Left^2, Right^2, Right^2, Right^2]$. We report all possible meaningful combinations in figure 3.4, but we spend more time discussing the *maximal* case where all of the foundations are enabled, and the two clusters point to different directions (see figure 3.5, BIBC).

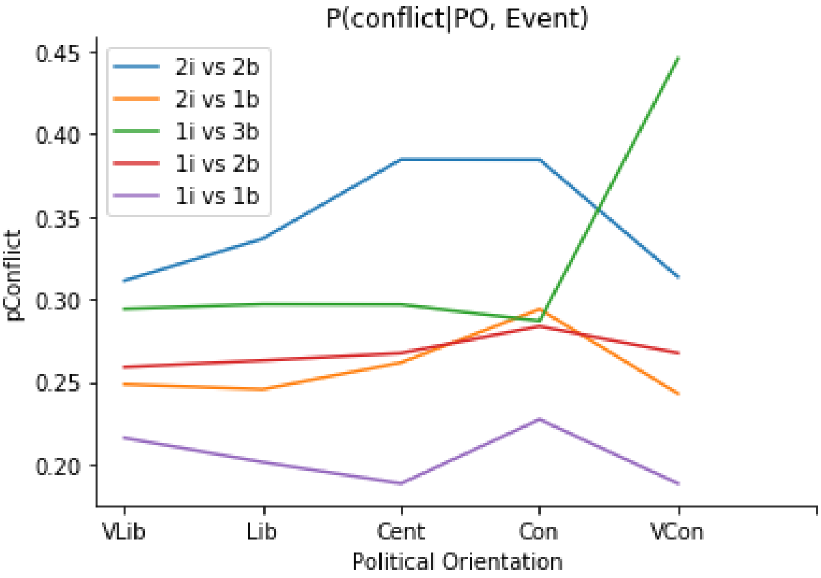


*Figure 3.4.* Probability of conflict over political orientation.

Another interesting pattern arising from the above results is that of the magnitude of conflict for each of the conflicting scenarios. We can see that BIBC triggers the greater probability of conflict, regardless political orientation, followed by WBC and then WIC. This is also confirmed by running diagnostic reasoning on the conflict variable, but this time having only observed the event's moral foundations leaving aside political orientation: 0.24, 0.26, 0.38, for WIC, WBC, and BIBC, respectively.

In summary, there are several interesting findings from PDM-A. First, we saw that on average conservatives are more likely to experience conflict than liberals, but this

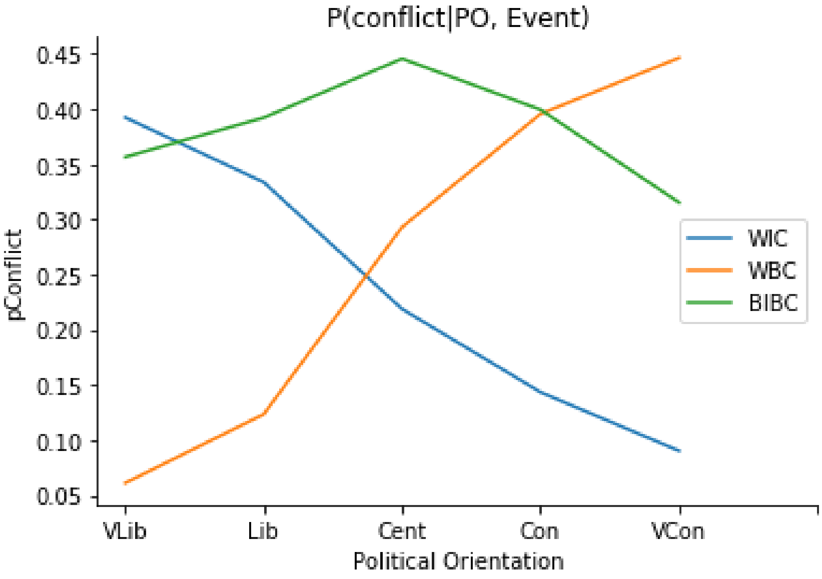


*Figure 3.5.* Probability of conflict over political orientation. *WIC*: Within Individualising Conflicting event. *WBC*: Within Binding Conflicting event. *BIBC*: Between Individualising and Binding foundations Conflicting event.

trend fades off for extreme conservatives who seem slightly less likely to experience conflict than conservatives (see figure 3.2). Considering responses to specific types of moral scenario, we see that liberals are more likely to experience conflict in a within individualising conflicting event, conservatives are more likely to experience conflict in a within binding conflicting event, and individuals who are on the centre of the political spectrum are slightly more likely to experience conflict in a between binding and individualising foundations conflicting event (see figure 3.5). Though interestingly, in this case as well, extreme liberals are more likely to experience conflict, and thus more probable to morally disengage, than extreme conservatives. Regarding the moral disengagement expected utility, for the most part this follows the same trend as the probability of conflict (see figure 3.3).

## 3.4 Discussion

The current study extends the conflict model presented in chapter 2 by addressing two gaps: 1) adding (preliminary) interactions between the agent's moral foundations so that the activation of one can influence the activation of another, and, 2) systematically exploring the ramifications of assuming that foundations coming from the same cluster tend to be more compatible than foundations coming from different clusters. Following these adjustments, the aim of the study remained the same: to explore how moral values might lead to conflicting beliefs, which in turn drive one to disengage from these beliefs. The findings were still in line with our main prediction: the farther one scores on the political spectrum, the more probable one is to experience conflict, and to disengage to resolve the conflict. This finding, for the most part, replicates the equivalent finding of the first version of the model (PDM), but this time the conflict intensity is slightly increased. In particular, in PDM the probability of conflict for extreme liberals, liberals, conservatives and extreme conservatives was 0.21, 0.21, 0.28, 0.27, while in PDM-A the equivalent values were 0.26, 0.28, 0.34, 0.34 (although, in actuality conservatives > extreme conservatives, but due to rounding they have the same value, the difference however is more apparent when looking figure 3.2). Adding interactions between agent's moral foundations leads to agents having more segregated moral matrices. That is, the difference between binding and individualising foundations increases. Having in mind the conflict function which states that the probability of conflict increases when two or more foundations are highly enabled and incongruent, we can explain the increase in conflict we observe in PDM-A. The difference between foundation clusters are now larger, and larger differences lead to larger disagreement, which in turn leads to inflated conflict values.

Noteworthy is the now smooth decrease in conflict from conservatives to extreme conservatives, in comparison to the steep drop found in PDM. That is the opposite pattern found in the more liberal political positions, where, initially a small rise spanning from very liberal to liberal was observed, but now this rise is steeper. That, in a sense, describes one aspect of the effects our latest adjustments had on the model results: the farther to the right we look, the more the increase added by the adjustments. Model-wise, that could be explained by the increased separation between clusters introduced by the added interactions in the agent's moral foundations. In other words, now that the activation of one foundation can influence the activation of another, we observe agents upholding moral foundations in a more segregated manner. For example, in PDM of the model it was possible to observe an agent having all five foundations highly enabled (although some of them might be silenced by event's moral foundations). This case now is less frequent, since the activation of the binding foundations will suppress the activation of the individualising ones (and vice versa to a lesser extent). Thus, allowing foundations interact with one another will result in agents having more discrete moral values in terms of foundation clustering, which will in turn lead to an increased chance of conflict for people who value the three binding foundations in a greater extend, flattening the drop from conservatives to very conservatives.

Regarding the conflict triggered within different conflicting scenarios, the findings again showed a similar pattern to these of PDM. When only the individualising foundations are enabled and conflicting (WIC), liberals are far more likely to experience conflict, and as a result disengage. On the contrary, when only binding foundations are enabled and conflicting (WBC), conservatives are the ones more probable to experience conflicting beliefs, pushing the agent to disengage more often. As PDM predicted, the fading off pattern of extreme conservatives was absent in WBC scenarios, as well as the fading off pattern of extreme liberals was absent in WIC scenarios. That is, extreme conservatives and extreme liberals are far more probable to experience conflict than conservatives and liberals in WBC and WIC scenarios, respectively. We also see that when the individualising foundations are turned off, the probability of conservatives to experience conflict decreases more than that of extreme conservatives. In other words, *care* and *fairness* foundations can explain at least some of the conflict probability in conservatives, while almost none in extreme conservatives due to the fact that conservatives value more equally all the five moral foundations than extreme conservatives, who seem to value the binding foundations significantly more than the individualising ones.

We found that left-wing political ideologies are more likely to experience conflict in a within individualising foundations conflicting event, while conservatives in a within binding conflicting event. However, there are differences in the probability of conflict each of these two political orientations experience under the different settings. In particular, extreme conservatives are far more likely to experience conflict in a within binding conflicting event than extreme liberals are in a within individualising conflicting event. This replicates the previous chapter's finding that conservatives have a higher probability to experience conflict even when we control for within individualising and binding foundations.

When we look at conflicting events triggered by *between* individualising and binding foundations (BIBC), things are again similar to what PDM predicted, with a few differences in the magnitude of the conflict. As in PDM, the results have a pyramid-like shape, peaked over the center of the political spectrum, and the probability of conflict dropping off at either end. The drop from centre position to either the left or the right of the political spectrum is substantially steeper than it is without the current modifications. In particular, the conflict values the PDM predicted for BIBC scenarios were 0.35, 0.37, 0.38, 0.37, 0.33, with a mean of 0.36 and a standard deviation of 0.02, while the same values of PDM-A are 0.36, 0.39, 0.45, 0.40, 0.32, with a mean of 0.38 and a standard deviation of 0.05. We should also discuss the increase in the mean conflict between the BIBC results of PDM and PDM-A, and more generally the increase in the conflict values. Technically, part of this rise can be attributed to the terms added in the conflict function. To remind you, we altered the conflict function so as to *intensify* the conflict based on the agreement of the event's moral foundations (i.e., whether or not the foundations belong in the same cluster). Now, since we only add to the conflict, rather than also penalise an opposing scenario (e.g., penalise the conflict when the foundations belong in the same cluster), it is reasonable to observe an inflammation in the average conflict value predicted by the model.

### 3.4.1 Known limitations

Although the current study works on a few potential limitation of the model presented in chapter 2, it does come with its own limitations. First, the choice of the parameter term in the conflict function is somewhat arbitrary. Intensifying the conflict coming from between clusters is not the only way to formally describe the hypothesis that beliefs coming from different clusters lead to an increased probability of conflict. For example, we could have also used a function which penalises the conflict coming from within clusters, or even both penalising and intensifying terms. Future research could try out these various options and see how conflict distribution changes. Using a function allowing within-cluster decrease in conflict and between cluster increase in conflict would probably render a different perspective to our question; segregation between clusters would potentially increase in a way that there would be a higher agreement within clusters and higher disagreement between clusters. This could lead to a decreased marginal probability of conflict, as within-cluster chance of conflict would decrease. We chose to to use only the intensifying term on the basis of the existence of conflict coming from the same clusters, but different options are available.

In the next chapter we present a study where empirical data where collected and compared against the model predictions. Besides validating the model, collecting empirical data also helps us decide between which of the two versions more accurately reflects reality. We saw that the differences between the older and the newer versions were limited only to the conflict and moral disengagement magnitude, but is the complexity introduced by the later updates justified? Does it worth to trade away some accuracy for some complexity? Collecting empirical data will help us answer these questions and validate our model.

To sum up, the current study advances the conflict model presented in chapter 2 by filling up a few theoretical gaps. In particular, PDM-A introduces interactions between the agent's moral foundations, and the concept of agreement between the event's moral foundations. With these, however, it also introduces some unavoidable complexity, which might not be justifiable on the grounds of advancement: the results of PDM-A are not far away from what PDM predicted. To clarify whether or not we should keep the alteration of PDM-A, in the next chapter we collect empirical data, and we compare it to model predictions.

# Chapter 4

# Moral disengagement and political orientation: Emprical study

## 4.1 Introduction

In the last two chapters we made an attempt towards what it seems to be the first formal model of cognitive processes that might trigger moral disengagement in a binary manner (i.e., either disengagement or not disengagement). We are aware that the literature introduces more shades of disengagement described as strategies (Bandura et al., 1996; Bandura, 1999) which occur embedded within both complex situational factors and cognitive processes. We chose to make a few simplifying assumption since the considerations of such variables as morality, political ideologies, and cognitive dissonance and disengagement, addressing them all in an interrelated manner, is a complex task on its own. Nevertheless, it is also an uncharted area whose results could lead to advances in the understanding of the processes of ideological polarisation (e.g., Neal, 2020), echo chambers (e.g., Garrett, 2009), political conflicts (e.g., Knutsen, 1995), and more. Complex analytical tools such as computational modelling, whilst powerful and insightful, they still need empirical data to validate them. The role of the current study is to compare empirical findings to model predictions in order to decide whether or not the theoretical model accurately reflects the reality.

### 4.1.1 Moral disengagement

A fruitful theoretical account regarding why people act unethically at work is moral disengagement (MD) (Bandura, 1986, 1999). Moral disengagement consists of a list of cognitive methods that enable a person to suppress their moral values and act immorally with no feeling (severe acute) of distress. Since the introduction of this theory, we have witnessed a growing research among almost every field in psychology and social sciences. In the field of individual differences, scientists have found that people's honesty-humility (Ogunfowora & Bourdage, 2014), leadership self-efficacy and affective motivation to lead (Hinrichs et al., 2012) authenticity (Knoll et al., 2016), interpersonal justice perceptions (A. Lee et al., 2019), perceptions of earnings management ethics (Beaudoin et al., 2015), moral identity (McFerran et al., 2010; Vitell et al., 2011; Kennedy et al., 2017), moral personality (McFerran et al., 2010), and religiosity (Vitell et al., 2011) all decrease the chance that persons will morally disengage. Furthermore, studies have established that resource exhaustion (K. Lee et al., 2016), mental entitlement (A. Lee et al., 2019), organisational identification (Chen et al., 2016), non calculative incentive to lead (Hinrichs et al., 2012), psychopathy (Stevens et al., 2012), and negative affects (Fida et al., 2015) all rise the chance that people will use one or more of the MD mechanism. Extending the MD theory (Bandura, 1986, 1999) R. A. Baron et al. (2015) discover that entrepreneurs' drive for profit was positively correlated with MD, whereas their drive for self-realisation had the opposite pattern. Employees' perception of mental agreement violation and work-related uncertainty are also positively correlated to MD (Astrove et al., 2015; Huang et al., 2017).

In several of these works, MD was a fundamental factor that explained the effect of individual differences on estimates of corrupt behaviour. During socialisation, individuals embrace moral values that function as mentors and as important foundations for self sanctions about ethical behaviour (Bandura, 1999). Throughout this procedure individuals supervise their behaviour in specific circumstances, judge it regarding their moral values and perceived conditions, and adjust it by the outcomes they apply themselves (Bandura, 2002). Therefore, hostility inhibitory mechanisms may occasionally get disregarded (Bandura et al., 2001; Bandura et al., 1996; Keltner & Robinson, 1996). Furthermore, extending sociocognitive theory, MD mediates the influences of self monitoring on immoral decision making (Ogunfowora et al., 2013), authenticity on immoral behaviour (Knoll et al., 2016), jealousy on social sabotage (Duffy et al., 2012), resource exhaustion on undermining (K. Lee et al., 2016), psychopathy on immoral decision making (Stevens et al., 2012), latent ideas on deception strategies (Tasa & Bell, 2017), observations of profits management ethics on morally questionable accounting exercises (Beaudoin et al., 2015), and employee creativity on workplace irregular behaviour (Zheng et al., 2019). Bringing together sociocognitive with attribution theories to describe why and when people conduct counterproductive work behaviour after they experienced mental agreement violation, Astrove et al. (2015) discovered that MD completely mediated the positive correlation between mental agreement violation and counterproductive work behaviour.

Although there are plenty of studies exploring the relationship between MD and various behavioural conduct, there is a research gap when it comes to MD and political ideology. To our knowledge, the relationship between ideology and MD has been investigated only by Jackson and Gaertner (2010), and Villegas de Posana et al. (2018). The scarce number of studies exploring this relationship only highlight their importance, and the lack of related evidence.

### 4.1.2 Ideology

Ideology has multiple definitions (Gerring, 1997); here, adopt a somewhat generic one suggested by Seliger (1976), who regarded ideology as a ”set of ideas ...[that]... explain and justify ends and means of organised social action and specifically political action” (p. 11). Later, an extension of this definition involved behaviour and beliefs in their concept of ideology, apart from ideas or values (Jost et al., 2009). Ideology can serve either as an explanation of the world (in the eyes of a particular person) or as a formula specifying how the world should be “. . . specifying acceptable means of attaining social, economic, and political ideals” (p. 309). That list of values, ideas, and behaviours has cognitive, affective, and motivating qualities, so that ideological dedication is a robust predictor of a variety of attitudes, choices, judgements and behaviours (Villegas de Posana et al., 2018). Moreover, ideology is often regarded as a system-justifying mechanism for rationalising how stuff are or needs to be (Zimmerman & Reyna, 2013). However, ideology is (at least) a bi dimensional construct, consisting of social and economic dimensions (Feldman & Johnston, 2014).

Jackson and Gaertner (2010) checked if people who scored high on the right-wing authoritarianism (RWA) scale, and people who scored high on the social dominance orientations (SDO) scale endorsed war as political interference and if such endorsement was mediated by MD. A different use of the mechanisms was predicted based on the ideology, but that prediction was not confirmed by the data. In actuality, either group utilised all disengagement mechanisms, mainly minimisation of consequences and moral justifications, but more strongly so RWA. Villegas de Posana et al. (2018), explored the usage of MD mechanisms by two Colombian illegal armed groups (*guerrillas*: leftist - revolutionary group (Wikipedia, 2020); *paramilitaries*: right-wing - national defence group), in addition to discrepancies among the groups. The results showed that the most frequently used mechanisms were: attribution of blame, euphemistic labelling, moral justification, and labelling with undesirable names (a type of dehumanisation). Moreover, the analysis uncovered differences among groups only in the amount of press releases, but not in the frequency or kind of the mechanism used. That is, either group used the same disengagement mechanisms and the rate of each was equivalent between the groups, a finding which is almost in line with what Jackson and Gaertner (2010) found: there are no significant differences between left and right ideologies in how frequent they use disengagement mechanisms.

Although the results and conclusions of the aforementioned studies almost point towards the same direction (i.e., the frequency of moral disengagement does not differ significantly among political orientations), broader research is needed since the existing one is limited to particular aspects of moral disengagement mechanisms. Specifically, both studies involve a war-like character either in the form of testing illegal armed groups (i.e., Villegas de Posana et al., 2018), or in the form of asking a single question regarding the approval of war as a political intervention(i.e., Jackson & Gaertner, 2010). These war-like elements of these studies might not engage the day-to-day disengagement mechanisms used by individuals in a typical day. Moreover, Villegas de Posana et al.'s (2018) sample (illegal armed groups) is not representative of the average political agent, making its result even less generalisable to the broader population. Research entailing more typical individuals and scenarios is thus needed.

### 4.1.3 Our study

The two studies mentioned above suggest that moral disengagement frequency and intensity is almost similar between different ideologies (Villegas de Posana et al., 2018) when the subject of disengagement entails violence. What about the more day-to-day events, and the more typical individual? Do conservatives disengage as often as liberals? Our model predicts that conservatives, overall, should disengage more than liberals. This is the question the current study tries to answer. An important note here is that, in order to make the results of the current study directly comparable to the predictions of the model we do not differentiate between the eight disengagement mechanisms, rather we tally out the different mechanisms into one general score of disengagement, according to the author of the disengagement scale (C. Moore et al., 2012). There is an issue this raises, to which we will return.

Briefly, do different ideologies tend to rely on different disengagement mechanisms more than others? Although the studies mentioned previously seems to suggest that they do not, we saw there are a few methodological issues: 1) the external validity of the conclusions (i.e., the extent to which the conclusions of these studies can be generalised to the wider population), as well as, 2) the validity of their measurement of ideology since they did not measure ideology directly. Furthermore, pre-existing literature in psychology (e.g., SDO, racism, etc) show that right-wing/conservative ideology is significantly more racist and accepting/endorsing of inequality and prejudice, which implies that such people should be more willing to use the moral disengagement mechanism *dehumanisation* than leftists In the same vein, one could use the same argument for right-wing associations, system justification, and victim blaming, which are more endorsed by conservatives. Thus, the scarce amount of studies (Villegas de Posana et al., 2018; Jackson & Gaertner, 2010) predicting no differences in disengagement frequency between political orientations, and the contrary implications posed by pre-existing research indicate that there is a gap on the topic. Our goal is to fill that gap, and thus contribute to our understanding of the psychological mechanisms behind experiencing conflicting beliefs and then disengaging from these beliefs. The frequency in the usage of disengagement directly implicates the frequency of cognitive dissonance/conflict experienced by people with differing ideologies. The second goal of the study is to gather empirical evidence to support or reject the predictions made by the conflict model presented in the previous chapters.

Based on our model predictions, we expect to find a relationship between political ideology and disengagement scores. Specifically, our model predicts that the more to the right one scores on the political spectrum, the higher their score on disengagement scale too. This relationship should break in extreme conservatives with them being less prone to use disengagement mechanisms than other rightists. We also expect that this relationship will hold true even after controlling for right-wing authoritarianism (RWA) and social dominance orientations (SDO), as it was found by the mediation analysis of Kugler et al. (2014): liberal-conservative difference in moral intuition was significantly mediated by RWA and SDO, in a way that conservatives' larger appreciation of ingroup, authority, and purity concerns can be attributed to greater levels of authoritarianism, while liberals' larger appreciation of fairness and harm avoidance can be attributed to smaller levels of social dominance.

It is worth mentioning at this point that for testing model predictions we merged two different datasets. The common questionnaires –and the ones we are more interested in– administered to participants in both datasets are the moral foundations and moral disengagement questionnaires, as well as a self-reported scale of political orientation. Evaluating model predictions only requires these three variables, although in one of the two datasets we also collected SDO and RWA data. Thus, the reader should bear in mind when reading the following sections, that the power of the predictions made about SDO and RWA is inferior to the rest of the results. We do elaborate on that when we discuss the limitations of this study in the discussion.

## 4.2 Methodology

### 4.2.1 Participants

This study has two sub-samples of data that were combined and analysed together, one coming from employees from various UK companies ($N = 276$, mean age $mu \simeq 30$, age range $18 - 60+$, females $= 84$). and the other consists mostly of Italian students of the Unisob University in Naples ($N = 64$, mean age $mu \simeq 30$, age range $18 - 60$, females $= 48$).

### 4.2.2 Measurements and procedure

The current study is a cross-sectional survey which measures the variables at interest in one single time. Participants had to fill an online questionnaires having four scales plus demographics. The order of the scales was randomised between participants while demographics were always completed at the end. The scales were: moral foundations questionnaire (MFQ), right-wing authoritarianism (RWA) social dominance orientations (SDO), moral disengagement (MDQ), and self-reported political orientation score. After the questionnaires participants filled a form of demographics.

**Moral disengagement questionnaire**

The MDQ (C. Moore et al., 2012) consists of 22 items in total, and eight sub-categories. Each item is on a 1-5 Likert scale where 1 is *strongly disagree* and 5 *strongly agree*. A total moral disengagement score can be obtained by averaging the scores within category and then summing these averages. In the current study we focus on the total score which indicates a general tendency to disengage. For the Italian sample we used the Italian, validated, version of the scale (Sili et al., 2014). Cronbach's alpha in our study 0.853.

### Moral foundations questionnaire

The MFQ (Haidt & Joseph, 2007; Graham et al., 2009) consists of two parts, 32 items, and six sub-categories (harm, fair, ingroup, authority, purity, and an extra one aiming to track participant's focus). In the first part, each item is on a 1-7 Likert scale with 1 indicating *not relevant at all* and 7 *extremely relevant.* In the second part, each item is on a 1-5 Likert scale with 1 indicating *strongly disagree* and 7 *strongly agree.* For scoring the questionnaire, we summed the scores within each sub-category, ending up having five (six with the focus category) scores, each representing how much a given participant upholds a given foundation. The Italian version was validated by BOBBIO et al. (2011). Cronbach's alpha in our study 0.861.

### Social dominance orientation

The SDO scale (Pratto et al., 1994) consists of 17 items each of which is on a 0-4 Likert scale where 0 is *completely disagree* and 4 is *completely agree.* For scoring the questionnaire, we summed all scores after reverse coding relevant items. The Italian version was validated by Roccato and Ricolfi (2005). Cronbach's alpha in our study 0.521. Note that this scale was only used in one of the two datasets.

### Right-wing authoritarianism

The RWA scale (R. A. Altemeyer & Altemeyer, 1981; B. Altemeyer, 2007) consists of 14 items each of which is on a 0-3 Likert scale where 0 is *completely disagree* and 3 is *completely agree.* For scoring the questionnaire, we summed all scores after reverse coding relevant items. The Italian version was validated by Roccato and Ricolfi (2005). Cronbach's alpha in our study 0.729. Note that this scale was only used in one of the two datasets.

### Self-reported political orientation

The self-reported political orientation questions is on a Likert scale from 1 to 7, where 1 indicated *very liberal* and 7 indicated *very conservative.*

#### Demographics

After filling the questionnaires, participants completed a demographic form consisting of their sex, age, political orientation (two times, once as a questionnaire item embedded within other previous items and once as demographic question; we did so as a measure of tracking participants' focus), and whether or not they were focused while they were filling the questionnaires.

### 4.2.3 Pre-analysis

#### Data exclusion

There will be three exclusion methods. First, we use the focus sub-category of the moral foundations questionnaire, where we exclude participants having a total score higher than five. The focus sub-category consists of two irrelevant questions which are just used to catch people who are not paying attention. For example, one of the questions is "When you decide whether something is right or wrong you take into consideration whether or not someone is good at math." Higher scores in this question indicate that the participant was not paying enough attention.

Second, if the participant answers *not focused at all* at the *how focused were you during answering the survey?* question, we exclude them from the analysis. We chose to analyse only data coming from participants who where partially or fully focused. Last, as another safety check we added two political orientation self-reported questions –one at the beginning and one at the end of online survey. If there is a discrepancy of more than three points between their answers, they are excluded. Again, this aims to test participants' focus. If for example, the first time they are asked they declare they are very conservative, while the second time they declare that they are very liberal (a difference of 4 in a 5-scores Likert scale) we assume that they were just not paying enough attention. We will also get advised from our regression models as for whether we should remove influential values. The way the online survey is structured does not allow for incomplete cases.

### 4.2.4 Analysis

After having removed outliers and created the political orientation variable, we run two linear regression models. First, to establish whether or not the tendency of moral disengagement can be predicted by political orientation, we run a simple linear regression with outcome variable *moral disengagement* and predictor *political orientation*. Now, in order to also test the unexpected model prediction presented in chapters 2 and 3, namely that the positive disengagement pattern from left to right breaks on very conservatives, we also run another linear regression but this time adding a quadratic term for political orientation predictor. We then compare the two models to decide whether or not our conflict model did a good job predicting that fading off pattern.

To evaluate whether the relationship between moral disengagement and political orientation holds true even after controlling for SDO and RWA, we incorporate the SDO and RWA variables into the above regression model. If the effect of political orientation on moral disengagement holds true after including these two variables then we can ask if the results matches model predictions, if it does not, then we could increase model performances if we added more components to the model's representation of political orientation. Last, we only have the SDO and RWA scores from the Italian sample, so we will only use this dataset when running this analysis.

## 4.3 Results

### 4.3.1 Descriptive statistics

In this sub-section we will present descriptive statistics for the two datasets both separately and merged. We will then discuss our results in relation to the descriptive statistics we present here. Table 4.1 shows some descriptive statistics for UK and Italian datasets separately. Table 4.2 includes descriptive statistics for both datasets merged.

Table 4.1
Descriptive statistics for each sample separately

| Variable | Sample | n | mean | sd | min | max | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|
| Harm | UK | 276 | 16.76 | 4.18 | 2 | 27 | -0.29 | 0.31 | 0.25 |
|  | IT | 64 | 24.02 | 2.94 | 15 | 30 | -0.63 | 0.13 | 0.37 |
| Fair | UK | 276 | 18.13 | 3.62 | 5 | 27 | -0.48 | 0.74 | 0.22 |
|  | IT | 64 | 23.55 | 2.51 | 18 | 30 | 0.01 | -0.40 | 0.31 |
| Loyalty | UK | 276 | 14.34 | 4.09 | 0 | 25 | -0.29 | 0.23 | 0.25 |
|  | IT | 64 | 19.50 | 4.59 | 3 | 30 | -0.78 | 1.55 | 0.57 |
| Authority | UK | 276 | 14.67 | 4.21 | 0 | 25 | -0.52 | 0.34 | 0.25 |
|  | IT | 64 | 16.78 | 5.01 | 2 | 27 | -0.65 | 0.22 | 0.63 |
| Sanctity | UK | 276 | 12.88 | 4.76 | 0 | 25 | -0.08 | -0.20 | 0.29 |
|  | IT | 64 | 13.97 | 5.99 | 2 | 24 | -0.29 | -0.89 | 0.75 |
| MD | UK | 276 | 18.04 | 6.63 | 8 | 37 | 0.48 | -0.45 | 0.40 |
|  | IT | 64 | 10.98 | 3.08 | 8 | 29.17 | 3.42 | 16.68 | 0.39 |
| RWA | UK |  |  |  |  | N/A |  |  |  |
|  | IT | 64 | 22.28 | 2.75 | 17 | 29 | 0.28 | -0.32 | 0.34 |
| SDO | UK |  |  |  |  | N/A |  |  |  |
|  | IT | 64 | 24.11 | 5.07 | 15 | 38 | 0.42 | -0.24 | 0.63 |
| Pol. Or. | UK | 276 | 3.17 | 0.73 | 1 | 5 | -0.38 | 0.47 | 0.04 |
|  | IT | 64 | 3.03 | 1.05 | 1 | 5 | -0.06 | -1.20 | 0.13 |

Table 4.2
Descriptive statistics for both samples merged

| Variable | n | mean | sd | min | max | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|
| Harm | 340 | 18.12 | 4.88 | 2 | 30 | -0.12 | -0.17 | 0.26 |
| Fair | 340 | 19.15 | 4.03 | 5 | 30 | -0.31 | 0.35 | 0.22 |
| Loyalty | 340 | 15.31 | 4.65 | 0 | 30 | -0.08 | 0.20 | 0.25 |
| Authority | 340 | 15.07 | 4.44 | 0 | 27 | -0.45 | 0.28 | 0.24 |
| Sanctity | 340 | 13.09 | 5.02 | 0 | 25 | -0.10 | -0.36 | 0.27 |
| MD | 340 | 16.71 | 6.71 | 8 | 37 | 0.69 | -0.33 | 0.36 |
| RWA | 64 | 22.28 | 2.75 | 17 | 29 | 0.28 | -0.32 | 0.34 |
| SDO | 64 | 24.11 | 5.07 | 15 | 38 | 0.42 | -0.24 | 0.63 |
| Pol. Or. | 340 | 3.14 | 0.80 | 1 | 5 | -0.33 | -0.02 | 0.04 |

### 4.3.2 Moral disengagement and political orientation

We first run a simple linear regression to test whether moral disengagement (MD) can be predicted by political orientation. The diagnostics of the first model indicated that there are five influential values and so we excluded them from the analysis and re-fit the model. The new model results showed that political orientation explained 2.1% of the moral disengagement variance, $F(2,333) = 7.27$, $p = 0.007$, $R^2 = 0.021$. The results showed a positive relationship between MD and political orientation, such that higher values of MD are linked to higher (more conservative/right-wing) values of political orientation, $b = 1.25$, $std = 0.46$, $t = 2.7$, $p = 0.007$, see also fig. 4.1).

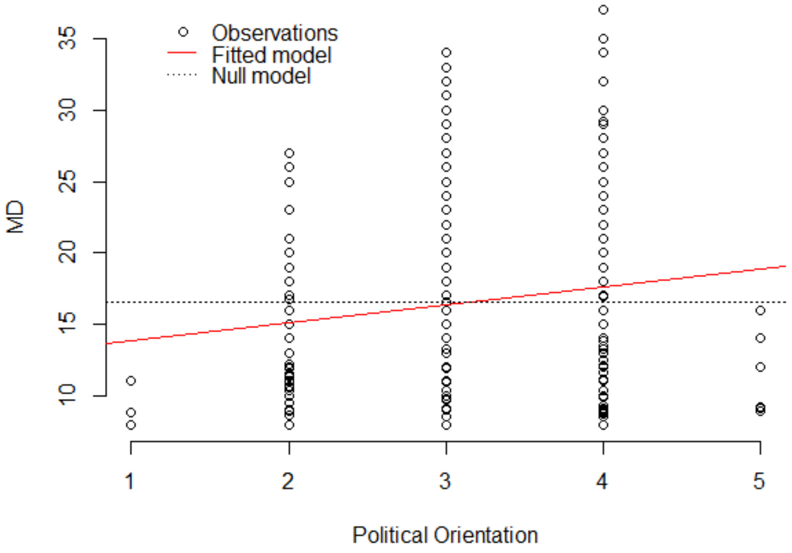


*Figure 4.1.* Moral disengagement predicted by political orientation: linear term

We established that there is a linear relationship between moral disengagement and political orientation, but the conflict model predicted otherwise. Specifically, the conflict model predicted that moral disengagement increases the furthest to the right one scores on the political spectrum, up until conservatives. This increasing pattern fades off, however, for extreme conservatives (see fig.2.3). To visually inspect if that is the case in our empirical data as well, we need to split our continuous political orientation variable to five categories (i.e., *very liberals* to *very conservatives.* Doing so allows us to more specifically test different bins of political orientation, and as the figure 4.2 depicts, although the model fit predicts a linear relationship between the two variables, the observations suggest a quadratic relationship.

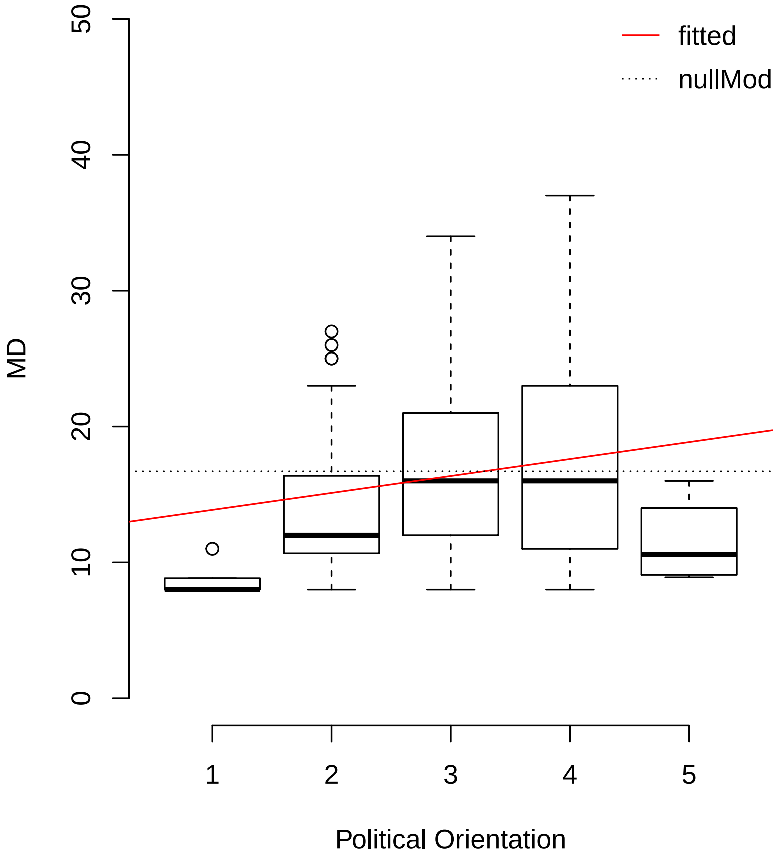


*Figure 4.2.* Moral disengagement predicted by political orientation: binned term

To test if a quadratic term offers a better fitting to the data, we run another linear model having this time a quadratic term representing political orientation. The model results $F(3, 332) = 12.41$, $p = 0.007$, showed that political orientation explained 7% of the moral disengagement variance $R^2 = 0.07$. The results showed that moral disengagement can be predicted by political orientation $b = 12.15$, $std = 2.67$, $t = 4.55$, $p < 0.001$, and that political orientation squared can predict moral disengagement over and above political orientation $b = -1.79$, $std = 0.43$, $t = -4.15$, $p < 0.001$ (see also fig. 4.3. Comparing the first model (linear term) with the second model (quadratic term), we get clear results. Both the *Bayesian information criterion* (BIC) and the *Akaike information criterion* (AIC) indicate that the linear model has a higher value (2226.673, 2215.231 respectively) than the quadratic model (2215.570, 2200.313 respectively).

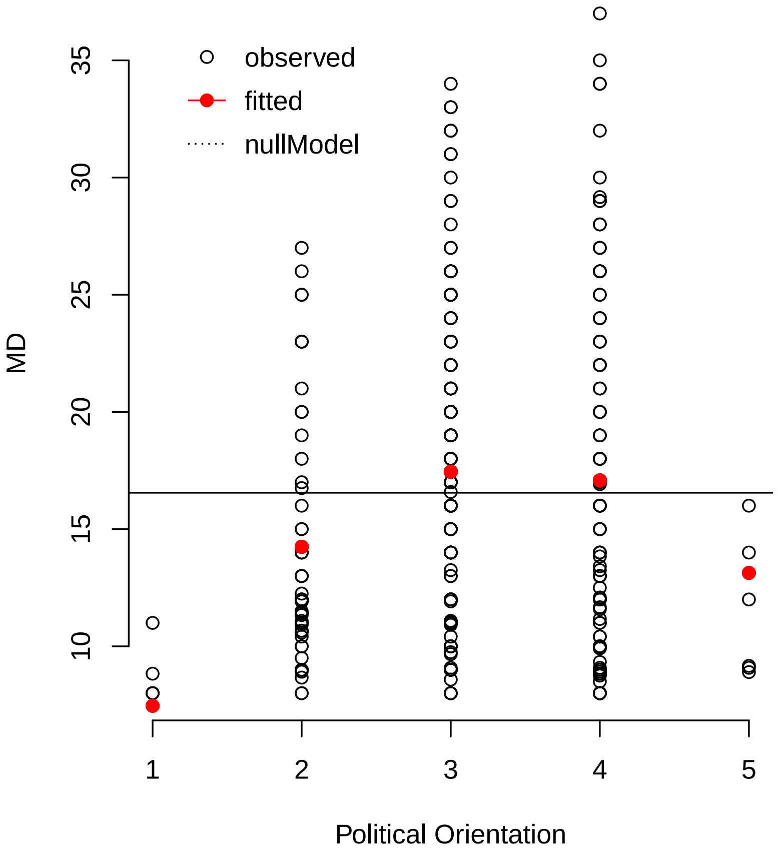


*Figure 4.3.* Moral disengagement predicted by political orientation: quadratic term

### 4.3.3 Controlling for RWA and SDO

In order to explore whether political orientation can predict moral disengagement over and above RWA and SDO, we run one last linear regression but this time adding RWA and SDO in the equation. The model results $F(4, 59) = 3.92$, $p = 0.007$, showed that the predictors explained 21% of the moral disengagement variance $R^2 = 0.21$. The results showed that only SDO can uniquely predict moral disengagement $b = 0.21$, $std = 0.07$, $t = 2.77$, $p = 0.007$, while the other predictors were not significant: political orientation

$b = 1.96$, $std = 1.13$, $t = 1.73$, 0.09, political orientation squared $b = -0.44$, $std = 0.25$, $t = -1.78$, 0.08, RWA $b = 0.24$, $std = 0.13$, $t = 1.79$, 0.079.

## 4.4 Discussion

The aim of the current study was to explore the variance in moral disengagement between different political ideologies. The trend of the conflict model predictions were for the most part accurate: the more right-wing a political agent is, the more the tendency to use disengagement mechanisms as a means to justify their morally questionable actions. This finding is inline with Jackson and Gaertner's (2010) results, where both RWA and SDO were positively associated with all moral-disengagement mechanisms,though more strongly so for RWA. On the contrary our results contradict Villegas de Posana et al.'s (2018) findings, where the average use of mechanisms did not differ between the two illegal armed groups, the leftiste Revolutionary Armed Forces of Colombia (FARC) and the right-wing (although, as we discussed in 4.1.2, the labelling left vs right may not be accurate) United Self-Defence Forces of Colombia (AUC), although there were minor differences regarding specific mechanisms. The differences between these studies might have to do with the varying methodologies, sample, and measurements used in each of them. Specifically, Villegas de Posana et al. (2018) use, as their sample, extreme, illegal groups which most certainly are different to some extend from our student- and employee-comprised samples.

The results of the current study support the prediction of the conflict model that the pattern of moral disengagement from conservatives to very conservatives drops. Specifically, both AIC and BIC criterion supported a fading-off pattern in conflict. Getting advised from the conflict model, this fading off pattern is explained probabilistically by noting that conservatives uphold more highly all five moral foundations while very conservatives seem to care more only for the binding foundations (loyalty, authority, purity). Endorsing more foundations could potentially lead to having a richer morality, which in turn may lead to having a higher chance of having conflicting beliefs and the need to decrease that conflict by disengaging. Very conservatives, on the other hand, only uphold three out of the five foundations, rendering them less probable to experience conflicting beliefs.

When right-wing authoritarianism and social dominance orientation effects are taken into account, the political orientation seem to lose its predictive validity. That is, the relationship political orientation has with moral disengagement is devoured by the relationship SDO and RWA have with moral disengagement. This could be because these two concepts are closely related to political orientation. In particular, it is well established by now that RWA has its intellectual roots in issues of fascism (Adorno et al., 1950) and is characterised by submission to perceived authorities, aggression when sanctioned by authorities, and adherence to social conventions (B. Altemeyer, 1994; R. A. Altemeyer & Altemeyer, 1996; B. Altemeyer, 1998). SDO is rooted in Social Dominance Theory (Pratto et al., 2006; Sidanius & Pratto, 2001) which argues that social groups are hierarchically arrayed in societies in terms of status and resources, and conflict among groups is minimised, in part, through socially shared beliefs that justify the hierarchy (e.g., legitimising myths). SDO, in particular, is an anti-egalitarian orientation characterised by a desire for hierarchical relations among groups (Pratto et al., 1994; Sidanius et al., 1994).

### 4.4.1 Limitations and future directions

In this subsection we will not only discuss the limitation of the current study, but also the limitations the current study has in relation to the conflict model presented in chapters 2 (PDM) and 3 (PDM-A). First, the conflict models compute the tendency of one to disengage as a function of conflict, while the current study measures directly the use of moral disengagement mechanisms. That is not so much a limitation as an observation: the conflict model gives a further insight as to how the tendency to morally disengage could come about. Specifically, the model suggests that moral disengagement takes place after having conflicting beliefs, while in the current study conflict is assumed.

On more limitation of our study is the way we validate the conflict model. Instead of directly feeding the empirical data to the model and test how probably model predictions are given this data, we compare model prediction and empirical result patterns. This way is enough to give us a qualitative *hint* as to whether model predictions are accurate, but not enough to quantitatively ensure that. Future studies could improve the model by incorporating a formal algorithm of testing a given dataset.

In the current study SDO and RWA are measured and used as control variables to test if the effects of political orientation on moral disengagement are robust to these aspects f conservative/right-wing ideology. These are not explicitly built into the conflict model, but rather implicitly subsumed by political orientation and moral values (Kugler et al., 2014; Lewis & Bates, 2011). However, this does not affect the comparability between them regarding the primary objective: to explore the tendency of moral disengagement based on political orientation. That is not affected by the absence of the SDO and RWA factors, though our results suggest that future development of the model may need to explicitly represent different aspects of ideology (i.e. not as uni-dimensional left-right, but rather as multivariate in nature).

Starting from there, following studies could extend the current project by increasing the sample size which will allow them to add SDO and RWA in the equation. Moreover, directly measuring conflict will make the empirical data more comparable to the conflict model, and the study will offer a more integrated answer of the question at hand.

Validating the conflict model builds up the credibility its other predictions have. Specifically, the model predicts that the most fertile ground for a conflict –be it within or between agents– is in a between individualising and binding foundations conflicting scenario. That finding could have implications for *echo chambers* and political *polarisation* research, where it is thought that people surround themselves with other people who have the same or relevant beliefs, in an attempt to avoid conflict, and reflect and reinforce their own beliefs (e.g., Garrett, 2009; Farrell, 2015). A current example of these mechanisms at play, is the recent segregation of the conservative end of the spectrum into Trump supporters and anti-Trumpers (called the Never Trump movement). The former group is deliberately isolating themselves by moving off of not only Fox News, but also Twitter and Facebook to Parler. In other words, they deliberately avoid conflict by creating their own echo chambers and reinforcing their own beliefs. As far as *conflict avoidance* goes, extending our findings to this line of research we would expect to see people creating "bubbles" based on the two clusters of foundations (individualising and binding), and that is because the biggest chance a scenario has to trigger a conflict, is when foundations between the two clusters are enabled and conflicting. In other words, in order to avoid conflict with other people, individuals who uphold individualising foundations will withdraw themselves from individuals who uphold binding foundations (and vice versa). This is a prediction which has been previously verified (see for example Baumann et al., 2020): liberals and conservatives tend to create their own echo chambers and "live in their own bubbles".

From a conflict avoidance perspective, individuals create their own environment with like-minded others to minimise exposure to conflict-eliciting information. Taking this a step further by looking at the conflict model findings, we expect liberals to form more dense/homogeneous echo chambers, meaning that they would be less probable to drift to alternative ideological viewpoints, than conservatives. This is because, according to our model predictions, conflict is *less* probable in a within individualising foundations scenario than it is in a within binding conflicting scenario. In other words, conservatives are more probable to clash with one another than liberals, leading to less dense/more fractionated echo chambers. As we saw in chapter 2, this theorising is partly supported by Eady et al.'s (2019), where 85 percent of liberals choose not to be exposed to an ideological opinion (Fox News) that more than 66% of extreme conservatives choose to. By contrast, 78% of conservatives choose not to be exposed to an ideological view (MSNBC) that 78% of liberals choose to. Therefore, whereas there is significant overlap in what both groups view, there are also zones of the media eco-system that are mainly seen by individuals of one ideological group, with conservatives being slightly more probable to expose themselves to different ideological viewpoint than liberals. There are certainly more factors that have a saying on this, but keeping everything else equal we would expect conservative echo chambers to be less dense since conflict is more probable in such environments. We investigate this prediction in chapter 5 were we present an agent-based model which explores how political agents position themselves based on conflict they experience with like-minded others.

### 4.4.2 Conclusions

In conclusion, the current study was the first of its kind and offered few significant insights as for how moral disengagement and political orientation are linked. Specifically, the main finding of this study was that right-wing political ideologies tend to use disengagement mechanisms more often than left-wing. The findings of the current stud were also used to validate the conflict model, presenting supporting evidence for the model predictions. The fading off pattern of moral disengagement from conservatives to very conservatives was only partially supported. The findings of this study could be seen as the first step towards exploring the frequency of conflicting beliefs within political agents, or to advance existing research on political polarisation and echo chambers.

# Chapter 5

# Political polarisation stemming from within ideology conflict: An agent-based approach

## 5.1 Introduction

Political polarisation has long been a central issue in political and social sciences, mostly with a focus on observing its progression throughout time, and investigating its causes. Time-wise, polarisation between U.S. Republicans and Democrats has monotonically increased since the mid 1970s (Abramowitz & Saunders, 2008); for the most part, technology has been targeted as being responsible. Specifically, rise in the number of users of various social media platforms, such as Facebook, Instagram, Linkedin, and others –and the time spent on them– has driven many researchers to investigate how users access, interpret, and evaluate social media content. Oftentimes this research is concerned with the concept of *echo chambers* (Sunstein, 2001a) or *filter bubbles* (Pariser, 2011), which postulates that political news engagement is highly fragmented due to polarisation (Lazarsfeld et al., 1944). In particular, echo chambers allow groups of people to strengthen their beliefs by interacting with others with whom they share similar views, and create filter bubbles which protect them from encountering contrary perspectives. Such detachment from, and ignorance of, alternative viewpoints is assumed to stem from a mixture of individual decisions (e.g., selecting which news pages to check or which accounts to befriend/follow), and the algorithmic forming of such decisions since search engines as well as social media platforms emphasise and suggest some sources over others (particularly to drive extended engagement, e.g., YouTube). As these algorithms get trained using individuals' decisions, and individuals take these decisions primarily from the suggestions promoted by the algorithms, a self reinforcing feedback loop gradually transforms decisions to a progressively limited and homogeneous list of options. As Bartlett (2015) commented about this *self-brainwashing* process where "certain ideas are repeated so often and with no contrary or alternative point of view that it fulfils the classic definition of brainwashing".

Although considerable effort has been devoted to studying echo chambers and political polarisation, rigorous theoretical questions for the exact operation of such processes between different political ideologies is sorely lacking. Most studies regard social media as the critical cause of echo chambers without giving much attention to individual factors such as morality and political orientation (but see also J. Baron & Jost, 2019). This line of research is difficult to reconcile with decades of audience research and given its glossing over of the psychology of users, is at least oversimplified and not particularly helpful for understanding the complex socio-psychological dynamics of public reception of, and connection to, news and media content. Thus, in the current study we use the validated predictions made by the conflict model presented in chapters 2 and 3, to develop an agent-based model which explores the phenomenon of echo chambers when the dissonance induced by conflicting beliefs is taken into account.

Investigating the psychological factors of echo chamber participants thus offers an important opportunity for understanding and potentially counteracting political polarisation. We expand the previous literature by introducing a new view on echo chambers formation within populations: agents not only select who they want to associate with based on agent-target similarity, but also based on the probability they have to clash with them. Specifically, we explore the emergence of polarisation as a result of just two rules:

1) similar agents attract each other, 2) conflicting agents repel each other. Complexity arises by embedding these rules in a landscape where moral conflicts cannot always be consistently resolved, even within a population containing otherwise homogeneous moral values/beliefs.

### 5.1.1 Political polarisation and echo chambers

Polarisation is a wide notion which focuses on the presence (or procedure of formation) of diverse groups differing on one or more attributes. Political polarisation, in particular, refers to distinction based on political ideology. Within each polarised cluster, it is usual to differentiate between mass polarisation, which appears across the electorate, and elite polarisation which appears across the elected. Here we focus on mass polarisation. This kind of polarisation is usually driven by ideologically narrow circles of like-minded peers, to the detriment of a detailed multi perspectival and evidence based comprehension of public affairs; a phenomenon called *echo chambers*.

Research on echo chambers follows on from the prolonged theory of selective exposure, that states that individuals prefer messages which are congruent to their beliefs while dodging incompatible views (Sears & Freedman, 1967). That is also supported by evidence showing that misinformation acceptance takes place when individuals select to pay attention and process information congruent with their beliefs (Kahan et al., 2012; A. Moore et al., 2021; Knobloch-Westerwick et al., 2020; Petersen et al., 2013). On the other hand, they might choose to seek flaws in the incoming (counter)information to reject it, even to the detriment of truth (Ditto and Lopez, 1992; but see also Lewandowsky and Oberauer, 2016)

Before the internet era –in which the accessible news sources were scarce– selective exposure in the pursuit of information did not generally emerge in circumstances of mass persuasion; however, post-internet people are able to reach (a huge amount of) information easier and modify what they want, thus they are more likely to select the content they will be exposed to (Tewksbury, 2005; Cass, 2007; Garrett, 2009). Online social networks have exacerbated this via their compound capability to enable users-news interaction in novel ways and to make use of complicated user-tracking algorithms to feed users with ideologically congenial content and drive ongoing engagement (Beam & Kosicki, 2014; Spohr, 2017). Such algorithms are constantly analysing user data and producing various sets of information displayed to each user, accordingly (Pariser, 2011). These algorithms (i.e., search engines) are widely used on the internet; for instance, Google's search engine uses such algorithm, as well as the news feed recommendation systems of online social networks such as Facebook (About-Facebook, 2019; Google, 2019).

That is not to say that echo chambers are an artefact of online social media. Online social media has intensified this phenomenon, but the desire to avoid challenges to core beliefs and opinions appears fundamental (Mutz and Martin, 2001; Sunstein, 2001b; Harmon-Jones et al., 2015; Festinger and Carlsmith, 1959; Festinger, 1962; although see also Garrett, 2009, for an opposing suggestion). Beyond core psychological drives/biases, echo chambers might also be fuelled by individual cognitive differences (Barkun, 2013). Specifically, data –coming from over one and a half thousands participants – on the Big Five, Right-Wing Authoritarianism (RWA), and demographics, suggest that age (positively), gender (lower in females), *openness* (positively), and RWA (negatively) predicted the amount of distinct news sources consumed. Individuals who consumed news solely offline scored higher in *Conscientiousness* and lower in *Neuroticism* compared to individuals who read the ”news feeds only” and the ones who read ”news feeds and offline”. Individuals who claimed they will not vote also stated the smaller number of distinct news sources consumed (Sindermann et al., 2020). To conclude, although online social networks might foster a fertile ground for the emergence of echo chambers, the actual causes have more to do with certain psychological factors.

In order to refine the need of psychological factors and diversity, recent simulation-based work has investigated echo chamber development in a virtual population of homogeneous rational (i.e., Bayesian) agents engaging in recurrent interaction (Madsen et al., 2017; Madsen et al., 2018). The results suggest an inherent propensity of social networks towards echo chamber development in spite of both an absence of cognitive differences among agents and their rationality. In other words, the structure of social networks per se seems sufficient for the formation of echo chambers (at least as instantiated by (Madsen et al., 2018). Other computational work (Fränken & Pilditch, 2020) also suggests that psychological biases and individual differences in the cognitive architecture of agents might not be necessary for the formation of echo chambers. The emergence of echo chambers was confined to the social agent population in which network-peers were selected on the basis of positive credibility estimates. In other words, when network-peers are selected at random and thus independent of whom one likes or perceives positively (asocial agents), no echo chambers emerge.

### 5.1.2 Agent-based modelling

There is a growing interest between social scientists in using system modelling tools to investigate how parts of a intricate phenomenon interact, persist or change, and ultimately understand the internal drivers of such systems. One such tool is agent-based modelling (ABM; e.g., Barrett et al., 2005; Yang et al., 2011)

ABM is capable of accommodating high diversity in agent traits and interactions between units and environment, along with features such as dynamics, feedbacks, and adaptations, which are unfeasible to capture with conventional statistical model approaches (Macy & Willer, 2002; Auchincloss & Garcia, 2015). Units can be described at various levels, including agent or group level (e.g., political groups, organisations, etc.). Scientific investigations that entail substantial diversity within and across units and various geospatial and relational components are well suited to ABM (Grimm & Railsback, 2005). In social science literature, simulations are used to investigate dynamic phenomena having various populations and surroundings like legal and health services, town bodies, individual citizens, families and more. Some ABMs involve thorough data and strive for realism (Barrett et al., 2005) while others are quite abstract (Yang et al., 2011; Axelrod, 1997).

The implementation of agents interactions can easily be governed by space networks, or a combination of structures (as highlighted in Alam and Geller, 2012). This would be far more complex to explain by mathematics, for example (Axtell, 2000). Significantly, agent-based models can regulate behaviours based on interactions at a specific distance and direction (thus allowing for action-at-a-distance). In addition, agent-based models also provide a robust and flexible framework for tuning the complexity of agents (i.e., their behaviours, degree of rationality, ability to learn and evolve, and rules of interactions). Another dimension of flexibility is the ability to adjust levels of description and aggregation. It is easy to experiment with single agents, sub-groups of agents, and aggregate agents, with different levels of description coexisting within a model. Thus, the agent-based approach can be used when the appropriate level of description or complexity is unknown and finding a suitable level requires exploration.

One crucial limitation of ABM is their computational limit introduced by their complexity. By their very definition, agent-based models consider systems at a disaggregated level, This level of detail includes the description of potentially many agent attributes and behaviours, and their interaction with an environment. The only way to treat this type of problem in agent computing is through multiple runs, systematically varying initial conditions or parameters in order to assess the robustness of results (Axtell, 2000). There is a practical upper limit to the size of the parameter space that can be checked for robustness, and this process can be computationally intensive, thus time consuming. Although computing power is increasing rapidly, the high computational requirement of ABM remains a limitation when modelling large systems (see Parry & Bithell, 2012).

A full discussion of ABMs and their characteristics is beyond the scope of the present contribution. In the remainder of this section we will therefore only refer to aspects of ABMs that are relevant to our model. Further details about ABMs and their general advantages can be found elsewhere (e.g., Madsen et al., 2019).

The three core elements that ABMs consist of are agents, patches, and links. Agents are the actors, and in our virtual world correspond to political agents. They are furnished with cognitive functions and possible behaviours, including attention and declaration (commit to a belief using a decision rule). All agents have the same cognitive functions and potential behaviours. Links represent connections between agents. In our work, the connections were signified by geographical contact, that is, two agents are bidirectionally linked to each other when they touch one another. Patches are building blocks of the environment in which agents act, although not so relevant to the current model, since, like agents, patches can change dynamically. They have been, however, used in other works to model, for example, fluctuations in fish stock at a specific location (Bailey et al., 2019).

Here, we reconstructed an idealised social network similar to Madsen et al. (2018) and Pilditch (2017) were agents form binary beliefs through interactions. As such, our agent-based model (ABM) captured agents' social environment and temporal dynamics of belief change, which both form necessary requirements for the observation of echo chambers.

### 5.1.3 Our study

Since there is considerable work done in echo chambers formation, where population architecture was taken into account, and since the results seem to be quite conclusive, here we focus more on two psychological variables: 1) the confirmation bias (i.e., the need to confirm our beliefs, usually by seeking out confirmatory information and/or like-minded others) (Lewicka, 1988, 1998; Klayman & Ha, 1987; Cosmides, 1989; Del Vicario et al., 2016; Ngampruetikorn & Stephens, 2016; Starnini et al., 2016) and, 2) cognitive dissonance and the need to decrease it as predicted by our conflict model in chapters 2 and 3.

Seminal work on confirmation bias (Schelling, 1969, 1971) showed that no firm incentives are required at a micro-level for group segregation to take place. In particular, even a weak individual preference to be around like-minded people is adequate for the creation of geographically isolated sub-populations and no high need to avoid dissimilar individuals is required. More research have reached to similar conclusions, demonstrating that even a sample of participants who strongly look for diversity may end up segregated (Zhang, 2004; Pancs & Vriend, 2007; Henry et al., 2011). All of these models portray agents that reside in the units of a grid or network and they have a distinct characteristic (e.g., skin colour, node shape etc) that commands their behaviour. They try to answer the question of how various micro-level interaction orders affect the macro-level (e.g., spatial distribution) of the agents. Importantly, the distinct characteristic is regarded as a rigid and immutable trait of the agents.

In our study, we first replicate the aforementioned results by developing a virtual population of agents with their only desire being to associate with like-minded others (a proxy for satisfying confirmation bias and reducing dissonance). Instead of enforcing and adjusting static network structures, we equip the agent dynamic connections so that collectives, and the variable social network where they exist, emerge and co-evolve. Subjects revise their belief (position on a 2-D grid) based on their neighbours; clusters develop and perish with varying dynamic asymmetries depending on whether agents link to similar or dissimilar others. These asymmetric connection dynamics include confirmation bias in the system and result in group separation. In the edge of extreme bias, we demonstrate that filter bubbles are somewhat unstable for conservatives while the broader group segregation (i.e., liberals) is stable and this finding persists in a broad range of differing parameters. We additionally demonstrate that firm confirmation bias typically raises the time the model needs to settle, and that poor bias can boost consensus formation, an interesting finding with implications for active social engineering.

Then, instead of incorporating the need to avoid dissimilar others, like some previous studies, we make the agents desire to avoid *similar* others. As we discussed in chapter 2, ideologically similar people can also conflict with each other. We visited a current example of conflict between ideologically similar groups, where the conservative end of the spectrum was segregated into Trump supporters and anti-Trumpers (called the Never Trump movement). In this example, Trumpers are deliberately isolating themselves by moving off of not only Fox News, but also Twitter and Facebook to Parler. That is, they deliberately avoid conflict by creating their own echo chambers and reinforcing their own beliefs. Our model will consist of agents having two opposing drives: 1) one that attracts them towards similar others, and 2) one that repulse them away from them. In that we are able to differentiate between liberals and conservatives who, as per our model results, have a different probability to clash with similar others. As far as *conflict avoidance* goes, extending our findings to this line of research we would expect to see people creating "bubbles" based on the two clusters of foundations (individualising and binding), and that is because the biggest chance a scenario has to trigger a conflict, is when foundations between the two clusters are enabled and conflicting. In particular, liberals have a lower probability to conflict with other liberals, –based on their moral beliefs– than conservatives have. If conservatives are more probable to experience conflicting beliefs with other conservatives, we expect to find their sub-population being less dense than liberals population. This study is the first that tackles segregation regarding *similarity* as a sedative factor for echo chambers formation.

## 5.2 Methodology

### 5.2.1 Model definition

The model is defined as follows. $N$ individuals, initially randomly distributed in a square box (which wraps both horizontally and vertically) of linear size $L$, –corresponding to a density $\rho = N/L^2$– execute a random walk of random direction and random number of steps between one to 10, with fixed step length $v \cdot \delta t$ and interact with the individuals they find within a certain distance $d = 1$ (agents located on the 8 neighbouring patches). Throughout the model, without loss of generality, we fix $\delta t = 1$. The position of agent $i$ at time $t$ is indicated as $[x_i(t), y_i(t)]$.

Each agent $i$ is characterised by two fixed state variables: 1) $p_i \in liberal, conservative$ reflecting the political orientation of the agent, and 2) $c_i \in (0, 1)$ representing their probability to experience conflict with like-minded others. Also, each agent is described by a dynamic binary state variable $h_i(t) = true, false$ reflecting whether or not the agent is happy surrounded by the given individuals. We use the conflict $c$[1] as a fixed parameter derived by the conflict model presented in chapters 2 and 3, while the $h$ internal state is

[1] Since in the conflict model political orientation was varying from 1 to 5 (extreme liberals to extreme conservatives), while in the current model political orientation is binary, we extracted the conflict pattern (conservatives more probably to experience conflict) rather than the exact values.

calculated at the end of each time point $t$. The $h_i$ variable is defined in a mood binomial space and takes into account the probability of conflict $c_i$ that agent $i$ has to clash with their like-minded neighbours, and the need $s_i$ to be surrounded by like-minded others. Upon interaction, agents modify their happy $h$ status seeking a local consensus with their neighbour so as to keep an equilibrium between conflicting with similar other and being surrounded by similar others. At each time $t$ the status of each agent $i$, $h_i(t)$, is updated as follows:

$$h_i(t) = C_i \leq r \wedge a_i(t) \geq (s_i n_i(t)) \tag{5.1}$$

where $C_i$ is the aggregated probability of agent $i$ to clash with either of their neighbours, $r$ is a random decimal number between 0 and 1 (inclusive) $r \in [0, 1]$, $a_i(t)$ is the number of alike neighbours of agent $i$ at time point $t$, and $n_i(t)$ is the agent-set of total number of neighbours surrounding agent $i$ at time point $t$ like so $n_i(t) = \{j : (x_i - x_j)^2 + (y_i - y_j)^2 = d_{ij}^2 < d^2\}$. Last, the *and* logical operation $\wedge$ indicates that both conditions (on the left and on the right of the operator) must be *true* in order for the outcome to be *true*. The aggregated conflict probability $C_i$ is calculated as follows:

$$\begin{aligned} C_i &= P(c_i(t)^1 \cup P(c_i(t)^2 \cup ...P(c_i(t)^k \\ &= \sum_{k,l} \left( P\Big(c_i(t)^k\Big) - P\Big(P(c_i(t)^k\Big) \cup P\Big(c_i(t)^l\Big)\Big) \prod_{k} \left( P\Big(c_i(t)^k\Big)\right) \end{aligned} \tag{5.2}$$

where $k, l = a_i(t)$. Put simply, equation 5.2 $C_i$ is the generic version of the union formula of probability: $A$ or $B$ or $C$ ... $k$, where $A, B, C$ are random events.

We can see from equation 5.1 that each agent strives for an equilibrium state between being surrounded by enough similar others to satisfy their need but no more than needed since each similar other increments the probability of the agent to experience conflict, and as per the conflict avoidance theory (e.g., Tjosvold & Sun, 2002; Barsky, 2011), people have the tendency to avoid conflicting environments. If each and every agent is happy $\{h(t)\} = true$ at a given time $t$, then the simulation reaches its equilibrium state and the iterations stop, a condition which is highly unlikely to happen since there is intrinsic randomness on whether each agent is happy (i.e., the probability to experience conflict). Heterogeneous groups are thus more fragile as individuals will tend to abandon them. Conversely, groups whose individuals experience the desired consensus will tend to persist in time with relatively stable mood status and position.

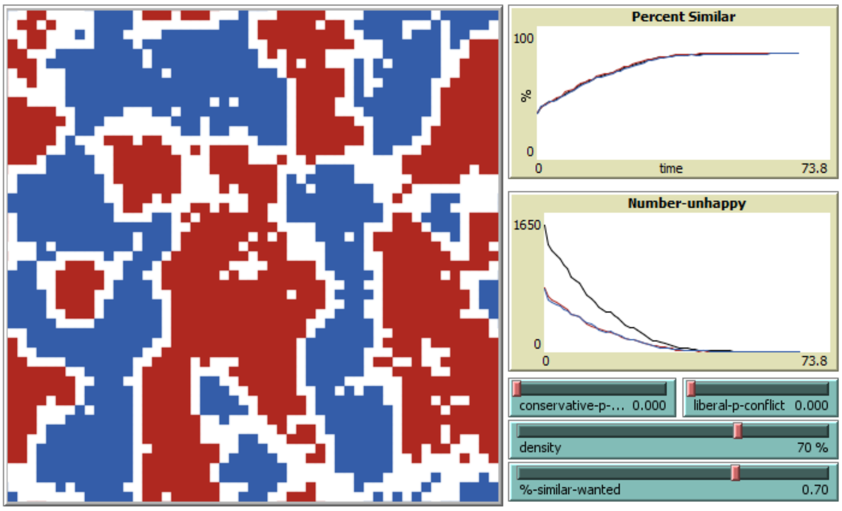
## 5.3 Results

The model dynamics are fully defined by four parameters: 1) the density $\rho = N/L^2$, 2) the probability of liberals to experience *within* ideology conflict $c_l$, 3) the probability of conservatives to experience *within* ideology conflict $c_c$, and 4) the percentage of similarity wanted $s$. The results presented here are obtained by numerical simulations of the model presented in the Methodology section.

### 5.3.1 Group formation

First we look into group formation setting liberals $c_l$ and conservative $c_c$ probability of conflict to zero, as a sanity check to see if our model can replicate the results reported by previous models. We set the density to 0.7 resulting in around 1820 political agents, $N = \rho * L^2$, where $L^2$ is the length of the grid, 51, squared. Having disabled the randomness introduced by the conflict parameters, the model can reach a local consensus based on just the proximity in the physical space, since in equation 5.1 the first term will always be true. The achievement of a local consensus favours the persistence of that proximity. As a result, when we have zeroed-out the conflict parameter, set the density to 0.7 and the similarity wanted $s$ to 0.7, the system reaches a stationary regime characterised by the presence of stable groups of individuals in around 60 iterations as can be seen from figure 5.1. On average, the similarity (total similar neighbours divided by total neighbours) of the two groups (abstractly, red: liberals, blue: conservatives) has equally stabilised at 80%. The results of this simulation, thus, replicate what older studies have found, namely that no strong incentives are required at a micro level for segregation to take place, just one variable (the desire to be surrounded by like-minded others) is adequate.

When adding the conflict variable , individuals can achieve a local consensus based on



conservative-p-... 0.000

liberal-p-conflict 0.000

density 70 %

%-similar-wanted 0.70

*Figure 5.1.* Group segregation with zeroed-out conflict variables

proximity in the social space, and reaching such consensus favours the persistence of that proximity. However, now the system reaches a quasi-stationary equilibrium characterised by the presence of meta-stable groups of individuals. One event can now change the regime of a group and change its composition and/or spatial properties, namely the arrival of a new individual and the spontaneous conflict by an agent. In the former case, either the newcomer's ideology (i.e., colour) is close to the group's local consensus and the newcomer will settle within the group, or they will leave if they do not belong to the same ideology as the rest members of the group. However, even if they are similar to the group, they might conflict with some members and leave. This dynamic interplay between the processes of group formation and group fragmentation introduces a rich phenomenology which can be understood in light of three quantities: the average fraction of moving agents, $N_m$, the average percentage of similar neighbours within each group $N_s$, and the average fraction of agents who were unhappy in $t-1$ and they are still unhappy in $t$, $N_u$. Formally:

$$N_m = (TN)^{-1} \sum_t N_m(t) \tag{5.3}$$

where $N_m(t)$ is the number of isolated and moving agents at time $t$,

$$N_s = S_n / T_n \tag{5.4}$$

where $S_n$ is the number of ideologically similar neighbours, and $T_n$ is the number of total neighbours, and last,

$$N_u = sum(ag(N_u(t)) + ag(N_u(t-1)) \iff 2)/ag(N_u(t)) \tag{5.5}$$

where $ag(N_u)$ is an agent-set vector of all the agents being either happy 1 or unhappy 0. The $\iff$ equality logical operator returns true if the left-hand and right-hand sides are the same.

Under these settings the model is highly unlikely to reach a stable regime since there is inherent randomness on whether the agent will experience conflict or not. Indeed, the model does not stabilises even after over 30.000 iterations. Therefore, we expect to observe some percentage of the agents to be unhappy, and a constant movement of agents around the grid. Groups are now less dense, with agents of one ideology drifting around to groups of the other ideology, in other words, $S_n$ is now smaller. In particular, conservatives seem to be slightly less surrounded by like-minded others, around 70%, than liberals, around 80%. That is because liberals have less incentive to break free from their ideology group, since they have a smaller probability to clash with one of the other (similar) agents. On the other hand, conservatives fluctuate more around groups since they tend conflict with like-minded others more often. The two forces –the need to be surrounded by similar others, and the potential conflict one might experience when surrounded by similar others– are more strongly suppressing each other in conservatives making them more *restless.* Figure 5.2 depicts model simulation. Note that now $x$'s indicate moving/unhappy agents.

The average fraction of moving/unhappy agents, $S_m$, also confirms the above finding. In each iteration conservative number of unhappy, and thus moving, agents is slightly but consistently higher than that of liberals. Again, this can be attributed to the higher probability conservatives have to clash with other agents similar to them. Having a higher number of moving agents leads to a smaller percentage of within group density and more a fluid group formation. On the contrary, liberals are more happy with their surroundings, having a smaller probability to conflict with one another, and so more stable regarding their group. In return, this leads to a higher group density and a more solid group formation.

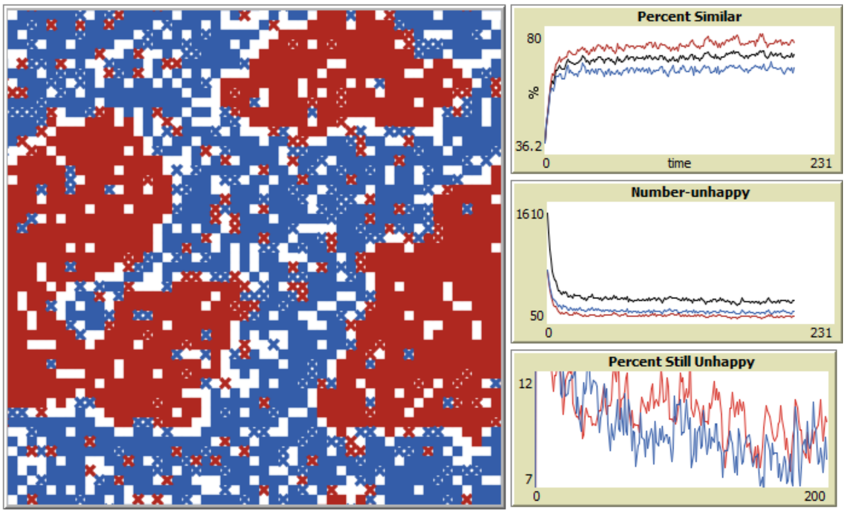


*Figure 5.2.* Group segregation with both conflict and the need to be surrounded by similar others variables

An interesting and unexpected finding concerns the percentage of unhappy/moving agents who where consistently unhappy for at least two iterations, $t-1, t$. We found that, although the percentage of unhappy liberals was smaller than the percentage of unhappy conservatives, a higher fraction of this percentage were still unhappy on the next iteration for liberals. Note here that, as equation 5.5 indicates, this percentage is calculated by dividing the number of still unhappy agents with the *total* agents of the same ideology, and not the number of unhappy agents. This is an interesting finding as it gives us some insights as to why the moving agents are moving. Specifically, there are two reasons one has to move: 1) need to belong to a group with like-minded others, and 2) conflict avoidance. Within an agent, conflict avoidance should be less consistent through iterations since it only comes about randomly, and only when one is surrounded by similar others, while the need to belong should always force someone to move as long as their current group does not consist of like-minded others. Therefore, we could safely assume that this emergent phenomenon of liberals being more consistently unhappy indicates that liberals' primary reason to move around is their need to be surrounded by similar others, while conservatives' primary reason is because they experience conflict with each other.

## 5.4 Discussion

This chapter studied the interplay between political orientation, conflict, and polarisation in the context of a simple model of echo chamber formation. The combination of these ingredients leads to the emergence of a meta-stable population structure in which groups of like-minded individuals spontaneously segregate in space, while single agents constantly leave or join them. Importantly, the emergence of the sub-group regime depends on two factors: 1) the need an agent has to be surrounded by like-minded others, and 2) the probability an agent has to clash with one of their surrounding like-minded others. Moreover, the sub-group structures are controlled, in terms of group sizes and stability, by the density of the agents. The feedback loop between mobility and conflict yields a strong assortativity between physical and ideology space: closer neighbours tend to share the same ideology. This scenario is transformed by the introduction of confirmation bias. The fact that individuals can be influenced only by peers sharing similar opinions leads to the emergence of echo chambers where polarised opinions coexist within the same group. Last, the results showed that a higher chance of clashing with like-minded others leads to more *restless* agents, and thus, less dense sub-groups. Specifically, conservatives, who, as the conflict model presented in chapters 2 and 3 predicted, are more probable to experience conflict with like-minded others, tend to form less dense echo chambers. That is in line with previous work that finds 85 percent of liberals are not being exposed to an ideological view (Fox News) that over two thirds of the most conservative quintile are being exposed to, while 78 percent of conservatives are not being exposed to an ideological viewpoint (MSNBC) that 78 percent of liberals are being exposed to (Eady et al., 2019). Thus while there is considerable overlap in information that both groups see, there are also areas of the media ecosystem that are primarily viewed by members of one ideological group, with conservatives being slightly more probable to expose themselves to a different ideological viewpoint than liberals.

Another, real-life example of conservative movement into different spaces to avoid conflict is conservatives' migration from Twitter/Facebook social media platforms to YouTube and Parler. After the 2016 presidential election the spread of fake news on social media became a public concern in the United States. To tackle this issue, online social media platforms introduced third-parties fact-checking programs which aim is to remove/reduce the spread of misinformation, and to inform users on the reliability of different posts/sources. At this point, Parler was founded and came out as "...the new Twitter for conservatives" (Aaron, 2010). Furthermore, relevant studies found that among those who shared any political content on Twitter during the election, fewer than 5% of people on the left or in the centre ever shared any fake news content, yet 11 and 21% of people on the right and extreme right did, respectively (Grinberg et al., 2019). Facebook and Twitter went as far as to ban accounts who were spreading misinformation, although this did not stop these users from spreading misinformation, it just prevented them from spreading misinformation on Twitter and Facebook. Now, online platforms like YouTube play a key role in the spread of conspiracy theories such as QAnon (Miller, 2021).

The contributions of the model are threefold. First, it shows that spatial segregation can result from a dynamics involving agents seeking consensus on ideology. Second, it provides a framework in which the sub-group structure, often assumed in the modelling of social systems, emerges from the microscopic rules of the model itself. Third, it shows that conflicting agents yield the possibility that different opinions coexist within the same sub-group. Furthermore, the current study is the first that explores echo chambers formation looking at the need people have to avoid *similar* others, rather than dissimilar others, which is extensively looked at by other studies (e.g., Del Vicario et al., 2016; Ngampruetikorn & Stephens, 2016; Starnini et al., 2016).

The parameter settings tested here have been chosen in an explorative manner. There may be other combinations of mechanisms that produce similar or even more realistic results. Future research could explore this possibility by employing a combinatorial optimisation algorithm such as simulated annealing to find optimal combinations of parameter settings in order to find the simplest relatively accurate model. Furthermore, future research could extend to multi-dimensional political orientation space: our model so far considers that political orientation is binary (i.e., either liberals/conservatives).

This one-dimensional description already captures how people might run off to bubbles consisting of like-minded others to avoid conflict, but in most real situations, political orientation comes with many variations. Hence, more complex mixed ideological space should be possible in an extension of our model.

It would be also interesting to investigate the characteristics resulting in the observed fractal patterns of real-life human space occupancy in archaeological records (d'Errico et al., 2012; d'Errico & Banks, 2013) or present-day distribution of cities (Arcaute et al., 2015; Arcaute et al., 2016). The research of the emergent network properties, such as the distribution and density of the sub-groups, and its evolution in time, would be interesting in light of the effect of such time-varying topology on spreading dynamics. At the same time, the emergence of sub-group structures where similar agents are relatively isolated from the rest of the sub-group is interesting with respect to the recently explored emergence of online echo chambers, where misinformation spreads and persists (Garrett, 2009; Bessi et al., 2015; Del Vicario et al., 2016; Fränken & Pilditch, 2020). Another interesting study, which relates to the spread of misinformation in and out of echo chambers, would be to explore how people update their beliefs when someone from their sub-group shares with them an opinion different from/congruent to theirs (Fränken et al., 2020). In the next chapter we present a study which explores how people revise their initial beliefs in light on new evidence coming from their friends.

To sum up, the current study used an agent-based model with simple rules to explore echo chamber formation/ideological polarisation. We found that just two rules are adequate for echo chamber to emerge: 1) the need an agent has to be surrounded by like-minded others, and 2) the probability an agent has to clash with one of their surrounding like-minded others. We also found that higher probability of conflict leads to less dense echo chambers and more mobile agents. We believe that the current work could open interesting possibilities for future research on the emergence of online echo chambers, in which misinformation spreads and holds. Equally interesting would be to investigate and compare real-life human space occupancy to the patterns we observed using with out agent-based model.

# Chapter 6

# Moral judgements revision in a social network: Modelling sensitivity to statistical dependencies in social learning

## 6.1 Introduction[1]

[2] Morality is usually one of the notable topics of infertile debates. An important reason for this is that moral judgements are resistant to change in light of new evidence (e.g., Lord et al., 1979; Skitka, 2010). Specifically, moral judgements are deeply entwined with emotion, motivation, and socialisation (e.g., Greene, 2008; Haidt, 2001; Prinz, 2007; Rai & Fiske, 2011) which is why being exposed to contradictory data does not necessarily change moral judgements. However, when new data comes in form of personal experience—instead of impersonal data—a single counterargument can suffice to induce a change in a moral judgements (Horne et al., 2015). For example, a single exposure to Pope Francis in combination with his appeal to counteract global climate

[1] OSF preregistrastion link: https://osf.io/rt6gs/

[2] This chapter is formatted as a CogSci publication submission that is why it has different format than the others.

change induced increased moral concern about climate change in adults from the United States (Schuldt et al., 2017). A different kind of evidence that has received less research attention is (online) social network evidence and potential dependencies underlying the judgements of social network peers. Do people revise their moral judgements when seeing what other people in their network believe, and if so, are they sensitive to dependencies that might exist between these judgements? Here, we address these questions formally, behaviourally testing predictions from a normative model across three conditions.

Recent computational models of belief revision which have focused on the non-moral domain represented judgements as scalars. One of the most prominent social learning models is the DeGroot model, where an agent's belief is represented as a point estimate which is revised by taking the weighted average of neighbouring judgements with fixed weights (see Becker et al., 2017; Banerjee et al., 2019, for recent applications of the DeGroot model). This implicitly makes the assumption that each neighbour's signal quality is the same, which usually does not hold in the real life (consider for example influencers on social media, which exert stronger influence on users than other users; (e.g., Lim et al., 2017), or simply that some members of a person's social circle may be more trusted, or more close, than others). Solutions to this limitation have been offered by Anunrojwong and Sothanaphan (2018) who see agents' beliefs as probability distribution over possible states of the world. In this way, signal quality can be communicated through the form of the distribution. For example, if an agent has belief μ and certainty (i.e., precision) τ indicating how certain they are about their belief, their belief can then be represented by a Gaussian N($\mu$, $1/\tau$). Although this proposal resolves limitations of the DeGroot model and its adaptations through a subjective quantifier of signal quality, it naively assumes that neighbours' judgements were statistically independent.

The assumption of independence is an important limitation of such previous work, since statistical dependencies naturally occur when confronted with information from sources of our social environment. Imagine, for example, that two friends told you they heard your favourite band will be playing in town this summer. Now, if one of your friends shared their knowledge with the other before both of them approach you, the amount of evidence provided to you is halved. This sequential belief updating process, which frequently takes place in real life, violates assumptions of independence (see Bikhchandani et al., 1992, for important theoretical work on sequential belief updating). Such violations have not only statistical relevance, but also real-life implications. We saw that when a group of people form their judgements based on the same or shared information, the amount of evidence shrinks. If individuals are unaware of such dependencies, or incapable of taking these dependencies into account when taking a decision, they might falsely allocate more credits on decisions coming from larger groups just because of the group size –usually observed during media influence, political campaigning, and conspiracy theories.

People are sensitive to such statistical dependencies in social learning, at least in some contexts. Specifically, using scalar representations of judgements, Whalen et al. (2018) showed that people revise their judgements about the number of blue vs. red marbles in an urn more when judgements of peers are independent as compared to shared or sequentially updated judgements. In Fränken et al. (2020), we extended this work using a normative framework including probabilistic representations of judgements. Precisely, Fränken et al. (2020) developed an experiment in which participants rated their prior belief on the suitability of two political candidates for public office using a dynamic (i.e., real-time updating) probability distribution. Thereafter, participants were exposed to the judgements of social network peers which varied in terms of their relationships with each other: independent peers having no relationship with each other; peers that formed their beliefs based on shared information; and peers that updated their judgements sequentially. Following exposure to peers' judgements, participants revised their initial judgements using the same interface. Statistically, the three condition were similar to Whalen et al. (2018), thus allowing to assess whether participants correctly weighed the evidential value of information between conditions. Results showed that participants distinguished independent from dependent sources of information (people are not simply combining their own judgements with the communicated judgements of their network neighbours. Rather, they are additionally sensitive to the origin of those judgements and to what extent they are redundant), but did not differentiate between the two dependent social network setups. These findings corroborated previous work on belief revision showing that people's beliefs are affected by the judgements of peers, and that statistical dependencies between sources of information influence the magnitude of belief updating.

Here, deploy two experiments and apply the above setup and interface to the domain of moral belief updating, which lacks recent computational attempts for modelling belief revision. Expanding previous work by Fränken et al. (2020), the present contribution investigates how people integrate *moral* information in a micro-social network. We first introduce a Bayesian framework to derive normative qualitative (directional) and quantitative predictions for how people should update their moral judgements. As in previous work (e.g., Whalen et al., 2018; Fränken et al., 2020), in experiment one, we compare three different conditions. The first serving as baseline: independent sources of information. In the second and third conditions, independence is violated: participants are either told that their peers formed their judgements based on the same data (shared information condition), or that peers communicated judgements prior to sharing them with the participant (sequential updating condition; see Fig. 6.1, for a summary). Importantly, in the sequential case, neighbour A communicates their judgements to neighbour B before both communicate their beliefs to the participant. In other words, optimally, only the information coming from neighbour B should be taken into account when the participant revises their judgements. Following a series of studies supporting that moral judgements are resistant to change (e.g., Lord et al., 1979; Skitka, 2010) we expected that participant would not update their judgements normatively as predicted by our model. Specifically, we expected that participants' posteriors will not be systematically different from their priors.

In experiment two, we slightly vary the experimental condition. Instead of a scenario where subjects learn what someone else (neighbours) thinks about the action of the main character (as in experiment 1), we now have a scenario where subjects judge the action and then learn about the motives of the actor from the sources (neighbours). In that way we can ensure whether they update their judgements at all –as predicted in experiment one–. Since the subjects will now be exposed to why the target behaved as they did, which will increase the credibility of the sources, we expect that participants will now update their judgements, but not normatively. In particular we predict that, We predict that our participants' patterns of judgments (in terms of both modal belief and certainty) will deviate from the normative model (cf Fränken et al., 2020). That means, the parameters we will extract from participants' reported judgements will not differ between conditions in the directions predicted by the model but may differ in magnitude in ways that we are yet to try to model. Put otherwise, we predict that there will be a negligible effect of network setups on participants' judgements i.e., people will not take into account informational dependencies when they are revising their judgements. Furthermore, we predict that prosocial motives will push participants' final judgement toward 'ethical', while selfish motives will move participants' final judgement toward 'unethical'. When the initial judgement of the participant is closer to 'unethical' in absolute terms (i.e. on the 'unethical' side of the midpoint of the belief rating scale), and prosocial motives push their final judgement toward 'ethical' in absolute terms (crossover response) we expect that shift in belief to be smaller than (a) what our model will predict and (b) the shift observed when judgements move in compatible directions (e.g. original ethical belief + prosocial motives will result in a larger update by comparison; same for unethical + selfish motives).

### 6.1.1 A normative framework

The normative framework models moral belief change from the perspective of a Bayesian agent across our three network setups. We initialise the model by computing subjects' moral beliefs which are based on reading about a fictional character facing a moral decision (i.e., asocial information). Subjects' prior beliefs are represented as beta probability distributions whose parameters are updated according to Bayes' rule. Here, the initial posterior probability of a belief or hypothesis $p(h)$, given asocial information, $d$, corresponds to the normalised product of the likelihood $p(d|h)$ and the prior $p(h)$:

$$p(h|d) \propto p(d|h)p(h) \tag{6.1}$$

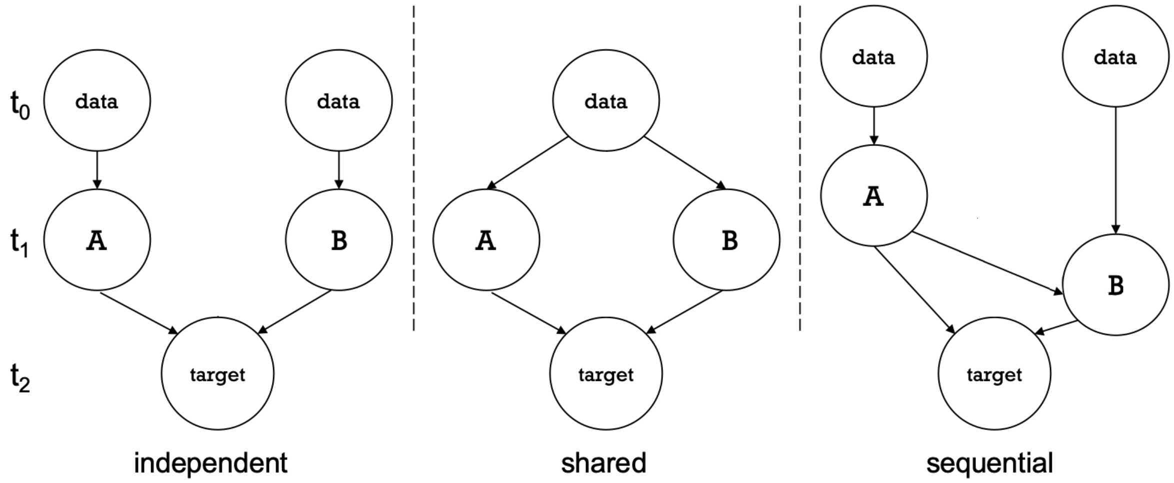


*Figure 6.1.* Illustration of network conditions. $t_0$: neighbors form beliefs given data. $t_1$: neighbors update beliefs based on interaction (sequential case only). $t_2$: target updates belief.

We choose a prior coming from a Beta distribution which is conjugate to a binomial likelihood $\binom{n}{k}p^k(1-p)k^{n-k}$ with $n = 2$ to ease modelling belief revision by using the analytical posterior Beta$(1+k, 1+n-k)$. For instance, from a uniform prior belief $X \sim$ Beta$(1,1)$, observing data [3] $D = \{1, 2\}$ where $k = 3$, the parameters of the posterior are $X \sim$ Beta$(2,3)$. This reflects the nature of participants' beliefs about the main character of the story, including a judgement about whether the character is ethical/unethical (mean of the beta distribution), and a measurement of certainty (precision of the beta distribution). Following the formation of priors based on asocial data, we derive qualitative (directional) and quantitative predictions for how people should normatively update their beliefs upon observing the beliefs of other network peers between the three experimental conditions. As in Equation 6.1, we use Bayes' theorem to model how people should integrate the beliefs (i.e., social information) from their network peers $s$:

$$p(h|s_1, ...s_n) \propto p(s_1, ...s_n|h)p(h) \tag{6.2}$$

In the case where the beliefs of neighbours are independent, the parameters of the posterior distribution of a participant's belief —after having been exposed to their neighbours beliefs— is equal to adding up the parameters of 1) the participant's prior distribution and

[3] In the present we do not quantify data, rather we assume that model prior is equal to people's prior plus some random noise. That is an arbitrary choice since in this study we are not interested in how people update their prior. We are interested in how people update their posterior after observing beliefs from peer.

2) neighbours' distributions (Fig 6.1, column 1). If participant's neighbours formed their beliefs based on the same information the model provides lower and upper bounds for the revised posterior. The parameters of neighbours are constant across conditions, and thus, any potential variation in updating can be attributed to manipulating the dependencies between neighbours. Therefore, the model lower bound is equal to subtracting the lowest $\alpha$ and $\beta$ parameters from the aggregate parameters. In the current study, we assume that all possible combinations of overlap are equally probable, and thus we model the normative impact of shared information on belief updating as having a magnitude intermediate between strictly independent information (higher magnitude) and sequentially updated beliefs (lower magnitude; Fig 6.1, column 2). Last, if neighbours updated their beliefs in a sequential manner, the model predicts that the parameters of participants' posterior distributions should be equal to adding up the parameters of 1) the participants' prior distributions and 2) the parameters of the neighbour who formed their beliefs last (i.e., Fig 6.1, column 3).

Given this framework, a normative agent will have a smaller difference between their prior- and posterior beliefs when beliefs of social network peers are dependent as compared to the independent. Moreover, our model predicts that in case of sequentially updated beliefs, a normative agent will update their prior beliefs to a smaller extend as compared to the case of shared information. Besides a few cases (e.g., Horne et al., 2015; Schuldt et al., 2017), however, moral beliefs have been shown to follow a somewhat stable trajectory over time (see also Williams Jr, 1979; Reuband, 1991), being resistant to change given new information (Lord et al., 1979; Skitka, 2010). We expect that the current setting will fail to predict people's responses. In particular, we do not expect to find any difference between people's beliefs before and after seeing what their neighbours responded.

## 6.2 Methodology

### 6.2.1 Participants

Participants (N = 69, range = 20 - 71, mean = 40 ± 12.6, 29 female) were recruited through Amazon's Mechanical Turk. Participants were native English speakers based in the United States. They were paid \$1.75 for their time ($\sim$ 20 minutes).

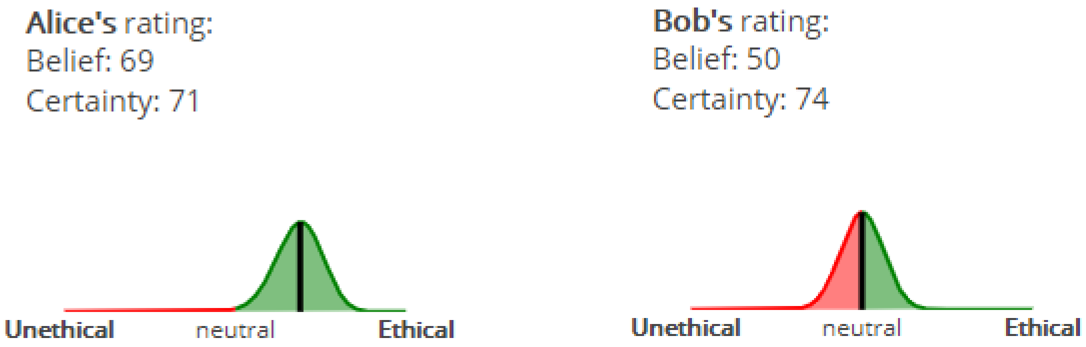
### 6.2.2 Task description and measures

In each trial, participants first read about a scenario in which a fictional character faced a moral decision (appendix A.4, experiment 1). After reading about the scenario, subjects provided their beliefs as for whether the fictional character is ethical or unethical. Following their initial belief ratings, subjects were exposed to the judgements of social network neighbours under consideration of statistical (in)dependence (see Fig. 6.2). Based on the judgements of their neighbours, subjects then revised their initial judgements to provide a final posterior judgement.

Before the main task, subjects filled a brief practising phase and a comprehension test to ascertain that they grasped how to supply their judgements via the interface shown in figure 6.3. In particular, participants provided their complete probabilistic judgements (i.e., beta densities) using two response sliders: one responsible for the density mean (belief slider), and the other responsible for the log precision (certainty slider). The range of the sliders was from one to 99; where a belief of 1 means the main character of the story is completely unethical, 50 is neutral, and 99 completely ethical. A certainty value of one is the weaker possible certainty, while a certainty of 99 is the strongest one regarding their judgements. The resulting density was updated in real-time and shown to participants as they selected their response. Visualisations were restricted to concave function shapes (i.e. $\alpha \geq 1$ or $\beta \geq 1$). Network neighbours were introduced as two potentially town-folks who read about a similar scenario as the participant to form their judgements about the main character of the story. After learning about their town-folk's judgements (and the underlying relationship between them), participants provided their posterior belief.

### 6.2.3 Design and Procedure

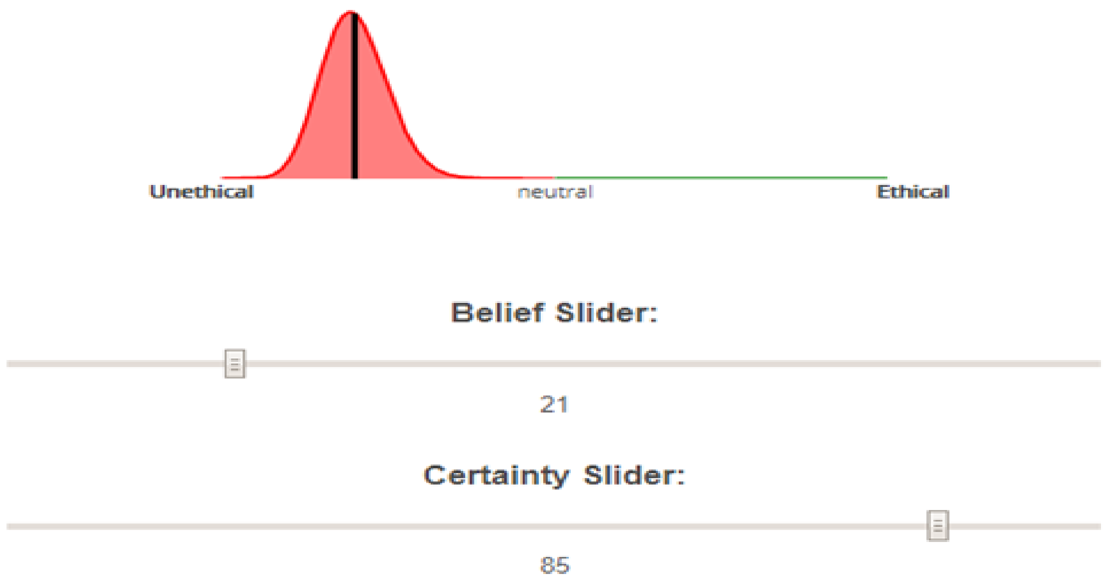
We deployed a within-subjects design with three levels: 1) independent information; 2) shared information; and 3) sequential belief updating. Since inferring model priors is beyond the scope of the present contribution, we computed normative priors based on participants' responses combined with Gaussian noise, $X \sim \mathcal{N}(0, 2)$. The normative prior and the parameters of the virtual neighbours are shown in Table 6.1. Parameter setup was constant among conditions, with the only source of variation being our independent variable (i.e., social-network set-up). Resulting model estimations and sufficient statistics are outlined in figure 6.4 (columns 1-2). The order of conditions and the scenarios were



**Alice's** rating:
Belief: 69
Certainty: 71

**Bob's** rating:
Belief: 50
Certainty: 74

*Figure 6.2.* Example beliefs of locals.



**Belief Slider:**
21

**Certainty Slider:**
85

*Figure 6.3.* Response interface

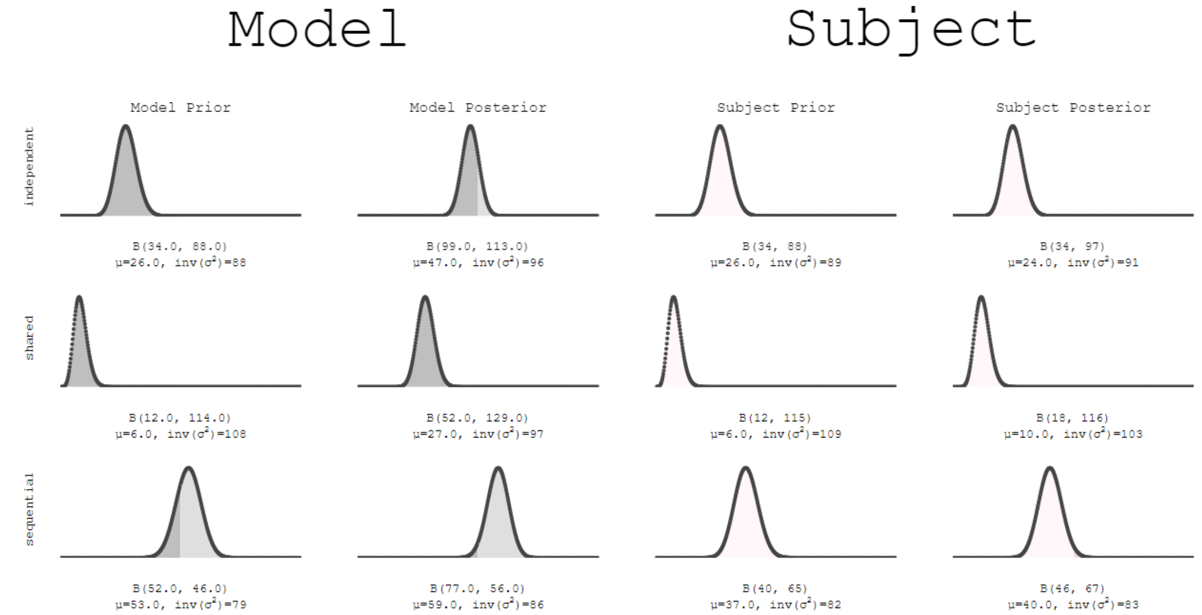


*Figure 6.4.* Summary of model predictions (columns 1-2) and behavioural results (columns 3-4) for each condition (rows 1-3). B($\alpha$ , $\beta$) refer to the aggregate parameters of model predictions / subject responses for each condition and measure (i.e. prior and posterior). In our analysis, we used the aggregate $\mu$ and $\sigma^2$ parameters (plotted below B($\alpha$ , $\beta$)) to make the interpretation of our predictions and results more intuitive.

randomised between participants. After the completion of the main task, participants provided basic demographics (e.g., gender, age, etc.)

Table 6.1
Fixed parameters used across conditions.

| Normative Prior | Parameters of Locals |
|---|---|
| subjects' prior + $\sim \mathcal{N}(\mu, \sigma^2)$ | B(80, 30), B(50, 20) |

### 6.2.4 Analysis

The analysis consisted of two sections. First, we aggregated the parameters of participants' posterior judgements withing condition and contrasted them with each other to assess qualitative (i.e., directional) alignment with the predictions of the normative framework. Therefore, we first compared participants' posterior means (m.1) and variances (m.2) among conditions using two linear mixed-effects models with condition (i.e. social network setting) as fixed effect and subject as random effects. Evaluating these parameters (i.e., mean and variance) independently could lose dependencies which might be present in how subjects revise these elements of their judgements. To address this, we calculated the Jensen-Shannon Divergence $D_{\mathrm{JS}}$ between prior and posterior distributions for each participant to establish if the degree of revising prior judgements varied by network set-up. $D_{\mathrm{JS}}$ enables evaluations of alterations both in mean and variance between distributions through a single symmetric distance measure given by:

$$D_{\mathrm{JS}}\left(P||Q\right) = \alpha D_{\mathrm{KL}}\left(P||Q\right) + (1-\alpha) D_{\mathrm{KL}}\left(Q||P\right) \tag{6.3}$$

where $D_{\mathrm{KL}}$ is the Kullback-Leibler Divergence, a common asymmetric measure in information theory for estimating how much a probability density $P$ has moved from a reference distribution $Q$. By definition, $D_{\mathrm{KL}} \geq 0$, is equal to 0 if and only if $P$ and $Q$ are identical. A limitation of $D_{\mathrm{KL}}$ is its non-symmetry, which is resolved by $D_{\mathrm{JS}}$ if $\alpha = 0.5$. Having computed each participant's $D_{\mathrm{JS}}$, we deployed one more mixed-effects model with condition as fixed effect and participants as random effects to contrast mean differences in the log-transform of $D_{\mathrm{JS}}$ (model 3). We used a log transformation of $D_{\mathrm{JS}}$ to address the residual distribution right skewness displayed by the non-transformed residual distribution of $D_{\mathrm{JS}}$. All models were compared to a reduced model having only an intercept as predictor and participants as random effects. Models were run in R with the function `lmer()` from the package `lme4` (Bates, 2010).

Then, we contrasted participants and model performances between conditions to test how far participants aligned quantitatively with the model predictions. Thus, we calculated $D_{\mathrm{JS}}$ between participants' posterior judgements (figure 6.4, column 4) and the normative posterior (figure 6.4, column 2) among conditions. Because of the skewed distribution of $D_{\mathrm{JS}}$ for this comparison, we deployed a Wilcoxon signed-rank test (non-parametric t-test; alternative hypothesis $> 0$) to test if subjects incorporated the asocial information as predicted by the normative framework. We also contrasted the distance between participants' and model posterior means, as well as the distance between participants' and model posterior variances (using two-sided unpaired t-test). We used the same analysis to evaluate whether participants update their judgements after being exposed to social information across our experimental conditions. In other words, we compared the prior mean/variance with the posterior mean/variance, and we also calculated the $D_{\mathrm{JS}}$ between prior and posterior distributions and compared that to zero.

## 6.3 Methodology

### 6.3.1 Participants

Participants (N = 123, age range = 18 - 65, mean = 34 ± 10.6, 40 female) were recruited through Amazon's Mechanical Turk. Participants were native English speakers based in the United States. They were paid \$1.75 for their time ($\sim$ 20 minutes).

### 6.3.2 Task description and measures

The task was quite similar to the first experiment. In each trial, participants first read about a scenario in which a fictional character took a decision (appendix A.4, experiment 2). At this point, participants do not know the motive behind the fictional character's decision. After reading about the scenario, subjects provided their judgements as for whether the fictional character is ethical or unethical. Without having access to the motive, we expect participant's responses to be neutral (neither ethical or unethical). Following their initial belief ratings, subjects were exposed to 1) the judgements of social network neighbours under consideration of statistical (in)dependence (see Fig. 6.2) and 2) the motive (pro-social/selfish) behind character's decision. Based on the judgements of their neighbours and the true motive, subjects revised their initial judgements to provide a final posterior judgement.

Prior to the main task, participants completed the same training phase as in experiment 1 (described in section 6.2.2), and a comprehension quiz to ensure that they understood how to provide their beliefs using the interface shown in Fig. 6.3.

Network neighbours were introduced as two potentially friends who knew about the main characters motives behind their decision. After learning about their friends' beliefs (and the underlying relationship between them), participants provided their posterior belief.

### 6.3.3 Design and Procedure

We used a withing-subjects repeated-measures design with two manipulated factors: 1) the three network setups (independent information; shared information; and sequential belief updating) 2) motive type (pro-social/selfish). The normative prior and the parameters of the virtual neighbours are shown in Table 6.1. Parameter settings were constant across conditions, with the only source of variation being our independent variable (i.e., social-network set-up). Resulting model predictions and sufficient statistics are summarised in Fig. 6.4 (columns 1-2). The order of conditions and the scenarios were randomised between participants. After the completion of the main task, participants provided basic demographics (e.g., gender, age, etc.)

### 6.3.4 Analysis

As with the first experiment analysis the analysis of the second experiment has two parts. First, we compared the aggregate parameters of subjects' posterior judgements between conditions to evaluate qualitative (i.e. directional) alignment with our model predictions. Thus, we first contrasted subjects' posterior mu (model 1) and var (model 2) between conditions using two linear mixed-effects models with condition (i.e. social network set-up) as fixed effect and subject as random effects.

In the analysis of the second experiment we also included the different motive types. Specifically, to evaluate if there are differences in belief revision based on motive type (pro-social/selfish) we run two more fixed-effects models. One with participants' posterior mu and one with participants' posterior var. The fixed effect structure for both models consisted by motive type, while the random effect structure consisted by each subject's intercept.

Finally, to evaluate whether participants' belief revision was significantly different than what the model predicted in the crossover condition (e.g., participant's initial judgement: ethical; final judgement: unethical), we run a Wilcoxon signed-rank test (non-parametric t-test) between participants' $D_{\text{JS}}$ (participants' difference between prior and posterior) and model $D_{\text{JS}}$ (model's difference between prior and posterior) focusing only in the crossover subset.

## 6.4 Results - experiment 1

### 6.4.1 Descriptive statistics

Table 6.2 presents the descriptive statistics for the main variables in our first experiment.

Table 6.2
Descriptive statistics for each condition

| Variable | n | mean | sd | median | min | max | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|
| Age | 69 | 40.03 | 12.69 | 39 | 20 | 71 | 0.61 | -0.33 | 1.52 |
| Pol.Or | 69 | 3.91 | 1.91 | 4 | 1 | 7 | 0.07 | -1.18 | 0.23 |
| **Independent condition** | | | | | | | | | |
| Mean prior | 69 | 0.41 | 0.14 | 0.44 | 0.04 | 0.72 | -0.34 | -0.26 | 0.02 |
| Variance prior | 69 | 88.42 | 9.5 | 88.86 | 50.49 | 126 | -0.15 | 4.91 | 1.14 |
| Mean posterior | 69 | 0.41 | 0.15 | 0.44 | 0.04 | 0.9 | 0.28 | 1.9 | 0.02 |
| Variance posterior | 69 | 88.73 | 9.44 | 88.98 | 48.59 | 126 | -0.34 | 6.09 | 1.14 |
| **Shared condition** | | | | | | | | | |
| Mean prior | 69 | 0.4 | 0.17 | 0.43 | 0.04 | 0.92 | -0.01 | 0.51 | 0.02 |
| Variance prior | 69 | 90.27 | 8.83 | 89.41 | 72 | 126 | 0.94 | 2.64 | 1.06 |
| Mean posterior | 69 | 0.40 | 0.17 | 0.43 | 0.04 | 0.92 | -0.10 | 0.57 | 0.02 |
| Variance posterior | 69 | 90.97 | 9.77 | 90.23 | 72 | 126 | 1.45 | 3.22 | 1.18 |
| **Sequential condition** | | | | | | | | | |
| Mean prior | 69 | 0.44 | 0.14 | 0.44 | 0.11 | 0.95 | 0.28 | 1.68 | 0.02 |
| Variance prior | 69 | 88.5 | 7.03 | 88.70 | 71.11 | 116.73 | 0.72 | 2.62 | 0.85 |
| Mean posterior | 69 | 0.43 | 0.14 | 0.43 | 0.12 | 0.89 | 0.46 | 1.51 | 0.02 |
| Variance posterior | 69 | 88.77 | 6.82 | 90.05 | 71.48 | 102.71 | -0.49 | -0.36 | 0.82 |

Pol. Or: Political orientation;

### 6.4.2 Sanity checks

Levene's test revealed that the homogeneity of variance assumption was maintained for all three dependent measures (all $ps > 0.05$) used between models 1-3. Inspection of residual plots confirmed that the residual posterior means, posterior variances and $D_{\mathrm{JS}}$ residuals were approximately normally distributed. Comparing each model to its reduced version revealed that inclusion of social network set-up as a factor did not explain a significant amount of variance in the outcome variable (see Table 6.3). Fig. 6.4 summarises model predictions and parameters of a random subject for our three experimental conditions.

Table 6.3
Overview of model fits for each variable (DV).

| Model | DV | $\text{BIC}_{\text{diff}}$ | $R^2_c$ | $\chi^2$ | $p$-value |
|---|---|---|---|---|---|
| 1 | $\mu$ | 9.58 | 0.06 | 1.08 | 0.582 |
| 2 | $\sigma^2$ | 7.24 | 0.15 | 3.42 | 0.181 |
| 3 | $\log D_{\mathbf{JS}}$ | 9.06 | 0.06 | 1.6 | 0.448 |

* $\text{BIC}_{\text{diff}} = \text{BIC}_{\text{full}}$ - $\text{BIC}_{\text{intercept-only}}$; $R^2_c$ = proportion of variance explained by the model. The p-value here is not that of the model, rather it indicates whether the the inclusion of the network setup predictor significantly improves model fit.

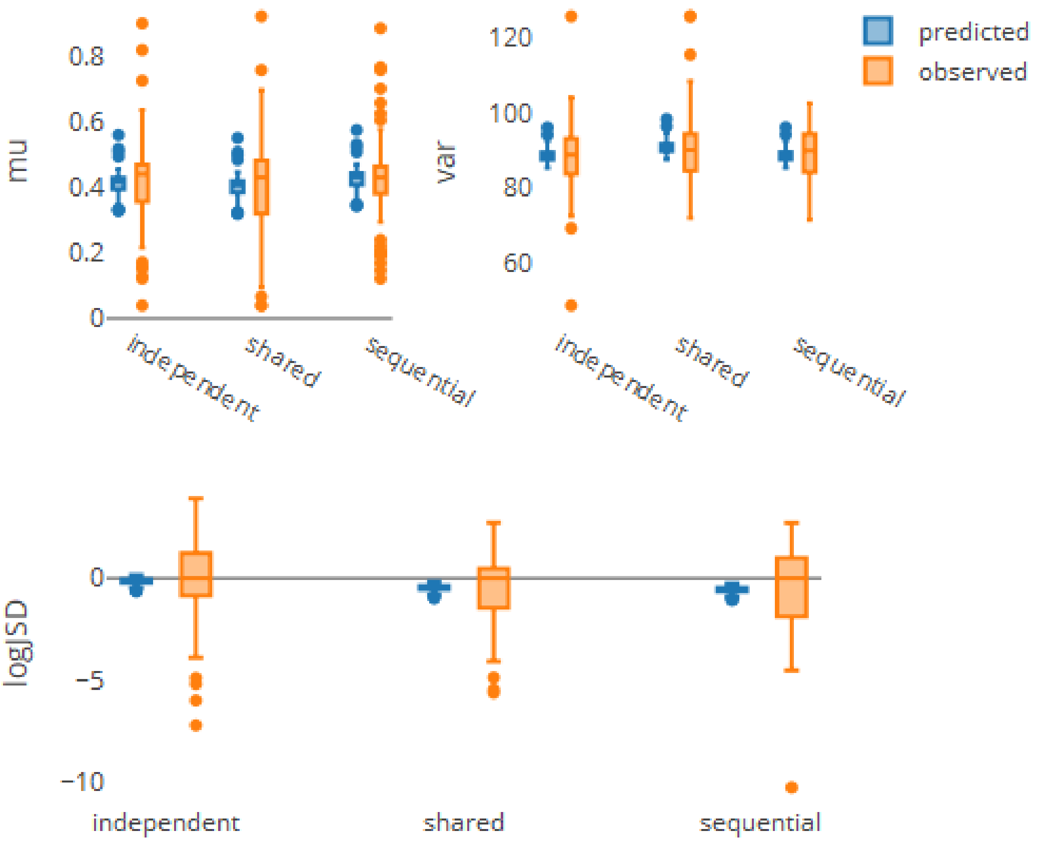


*Figure 6.5.* Boxplots for observed and predicted values for the three models of the qualitative analysis. In all three cases there were no differences among the three network setups.

### Qualitative comparison

Overall, the results of the three linear mixed-effects models below point towards the same direction: the distributions of participants' responses across network setups do not differ systematically from each other (Fig. 6.5).

To test whether the mu parameter of participants' response distributions can be predicted by the different network setups (i.e., independent/shared/sequential), we ran a linear mixed-effects model, with network setups as fixed effects and random intercepts for subjects. The overall model fit was not significant, $R^2(R^2_c) = 0.059$, and the marginal exploratory power of the social network factor was $R^2(R^2_m) = 0.007$, $F(2, 244) = 0.8$, $p > .05$. In brief, there were no differences on the mu parameters of participants' response distributions across network setups.

We repeated the above analysis to test whether social network setup affected the var parameter of participants' responses. The overall model fit was not significant, $R^2(R^2_c) = 0.64$, and the marginal exploratory power of the social network factor was $R^2(R^2_m) = 0.002$, $F(2, 244) = 1.14$, $p > .05$. In brief, there were no differences on the var parameters of participants' response distributions across network setups.

Finally, we repeat the previous analyses using the log Jensen-Shannon Divergence ($D_{\text{JS}}$) of participants' response distributions as the dependent variable. The overall model fit was not significant, $R^2(R^2_c) = 0.12$, and the marginal exploratory power of the social network factor was $R^2(R^2_m) = 0.001$, $F(2, 244) = 0.18$, $p > .05$. In brief, there were no differences on the $D_{\text{JS}}$ of participants' response distributions across network setups.

### Quantitative comparison - Predictions vs Observed

Taken together, the results of the quantitative comparisons below suggest that the model predictions and participants' responses were different, and the difference seems to be driven by participants being more confident than what the model predicted.

We conducted an unpaired t-test between participants' posterior mu's and model posterior mu's to determine if the mean of the participants' posterior distributions was different than that of the model's posterior. The results showed that there was no difference between the mu parameters, $t(393.35) = -1.242$, $p = 0.215$, with participants mean mu being 0.42, while model prediction 0.43 (confidence interval of the difference $[-0.04, 0.01]$). That is, the difference between participants' mu parameters and model mu parameters is close to zero.

To check whether the var parameter of participants' posterior distribution was systematically different than the var of model posteriors, we run an unpaired t-test between participants' posterior var's and model posterior var's. The results showed that there was a significant difference between the var parameters, $t(262.87) = 7.98$, $p < .001$, with participants mean var being 0.0017, while model prediction 0.0012 (confidence interval of the difference $[4e - 04, 6e - 04]$). That is, the difference between participants' var parameters and model var parameters significant, with participants' var being less than what model predicted.

In order to evaluate whether the $D_{\text{JS}}$ between participants' and model posteriors was greater than zero, we run a Wilcoxon signed-rank test (non-parametric t-test) using as alternative hypothesis $> 0$ and feeding it the distribution of $D_{\text{JS}}$. The results showed that the $D_{\text{JS}}$ distribution was reliably different than zero, $V = 14430$, $p < .001$. That is, the difference between participants' response distributions and model prediction is greater than zero, which implies that participants' responses were different than model predictions.

### 6.4.3 Quantitative comparison - Participants priors vs Participants posteriors

In order to evaluate whether there was a difference between the mu, var, and $D_{\text{JS}}$ metrics we run three t-tests (paired, paired, Wilcoxon signed-rank). The results of all three tests were non-significant: $t(206) = 0.096$, $p = 0.923$, $[-0.0126, 0.0139]$; $t(206) = 0.705$, $p = 0.482$, $[0, 1e - 04]$; $V = 14430$, $p > .05$; respectively. Overall, the results of the final comparisons seem to suggest that participants' prior judgements were not different than their posterior judgements. Figure 6.6 shows the results of the quantitative comparisons.

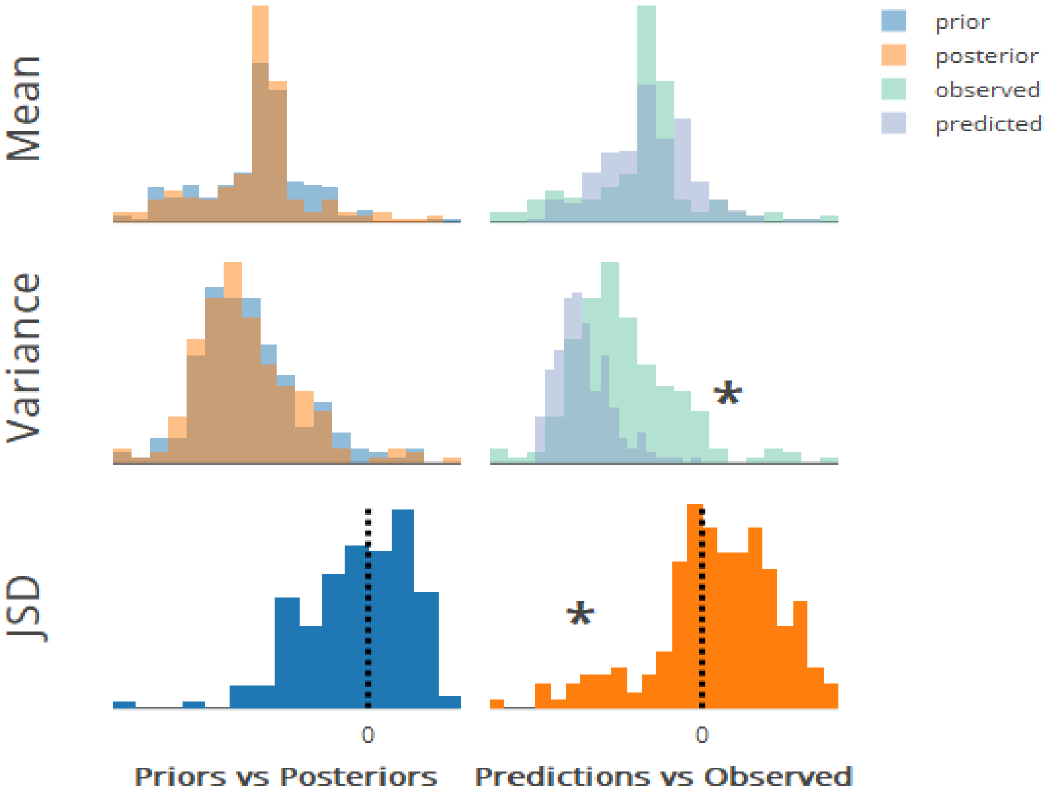


*Figure 6.6.* Results of the quantitative analysis. The left column depicts the comparison between subjects' priors and posteriors, while the right column depicts the comparison between predicted and observed posteriors. The rows outline the three different distribution metrics: mean, variance, and log Djs. In the bottom row, the null hypothesis is if the distribution is different than zero. * indicates the presence of a significant difference.

## 6.5 Results - experiment 2

### 6.5.1 Descriptive statistics

Table 6.4 presents the descriptive statistics for the main variables in our first experiment.

Table 6.4
Descriptive statistics for each condition

| Variable | n | mean | sd | median | min | max | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|
| Age | 123 | 33.46 | 10.63 | 30 | 18 | 65 | 1.26 | 0.88 | 0.95 |
| Pol.Or | 123 | 3.88 | 1.87 | 4 | 1 | 7 | 0.08 | -1.06 | 0.17 |
| **Independent condition** | | | | | | | | | |
| Mean prior | 123 | 55.14 | 20.65 | 55 | 5 | 95 | -0.30 | -0.27 | 1.86 |
| Variance prior | 123 | 52.24 | 26.55 | 58 | 1 | 99 | -0.19 | -0.88 | 2.39 |
| Mean posterior | 123 | 0.57 | 0.23 | 0.6 | 0.05 | 0.95 | -0.48 | -0.4 | 0.02 |
| Variance posterior | 123 | 0.01 | 0.02 | 0.01 | 0.001 | 0.08 | 2.22 | 4.33 | 0.01 |
| **Shared condition** | | | | | | | | | |
| Mean prior | 123 | 55.55 | 19.16 | 55 | 5 | 95 | -0.38 | 0.23 | 1.73 |
| Variance prior | 123 | 47.73 | 26.55 | 50 | 1 | 99 | 0.02 | -1 | 2.39 |
| Mean posterior | 123 | 0.56 | 0.18 | 0.6 | 0.05 | 0.95 | -0.36 | -0.42 | 0.02 |
| Variance posterior | 123 | 0.01 | 0.02 | 0.01 | 0.001 | 0.08 | 2.05 | 3.36 | 0.002 |
| **Sequential condition** | | | | | | | | | |
| Mean prior | 123 | 55.62 | 17.3 | 51 | 5 | 95 | -0.07 | 0.17 | 1.56 |
| Variance prior | 123 | 43.01 | 24.54 | 48 | 1 | 99 | 0.08 | -0.71 | 2.21 |
| Mean posterior | 123 | 0.55 | 0.21 | 0.55 | 0.05 | 0.95 | -0.34 | -0.63 | 0.02 |
| Variance posterior | 123 | 0.01 | 0.02 | 0.01 | 0.001 | 0.08 | 2.03 | 3.24 | 0.002 |

Pol. Or: Political orientation;

### 6.5.2 Sanity checks

Levene's test revealed that the homogeneity of variance assumption was maintained for all five dependent measures but one (all $p's > 0.05$ except means motive type model with $p = 0.002$) used between models 1-5. Inspection of residual plots confirmed that the residual posterior means, posterior variances and $D_{\mathrm{JS}}$ residuals were approximately normally distributed.

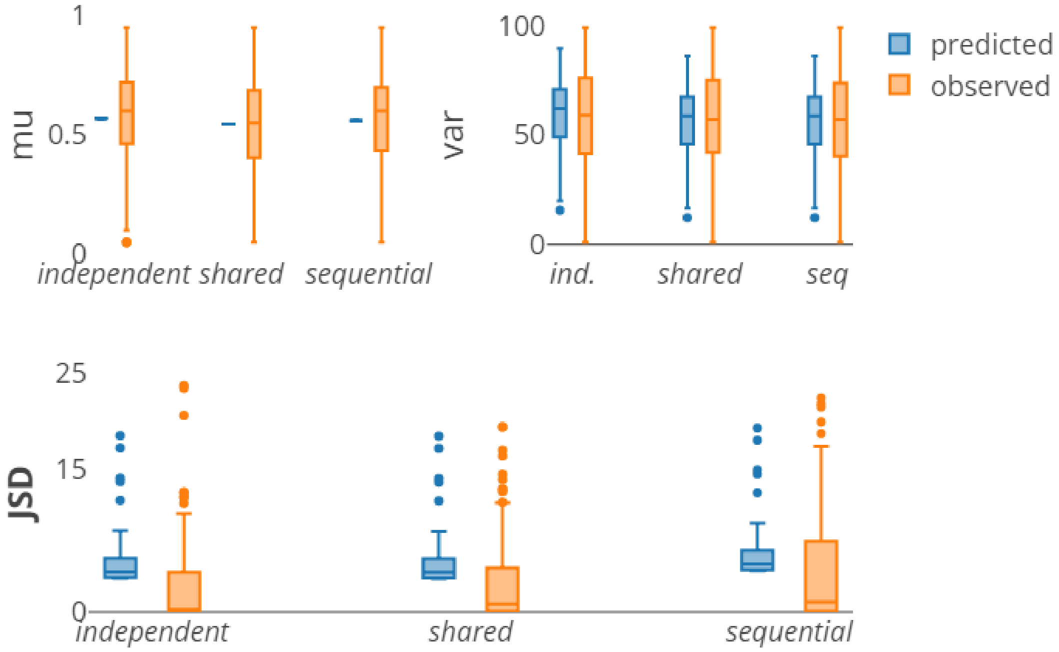


*Figure 6.7.* Boxplots for observed and predicted values for the three models of the qualitative analysis. In all three cases there were no differences among the three network setups.

### Qualitative comparison

Overall, the results of the three linear mixed-effects models below point towards the same direction: the distributions of participants' responses across network setups do not differ systematically from each other. In other words, participants responses were similar between conditions (Fig. 6.7).

To test whether the mu parameter of participants' response distributions can be predicted by the different network setups (i.e., independent/shared/sequential), we run a linear mixed-effects model, with network setups as fixed effects. The random effect consisted of an intercept for each subject: $(1|subject)$. The overall model fit was not significant, $R^2(R^2_c) = 0.022$, and the marginal exploratory power of the social network factor was $R^2(R^2_m) = 0.002$, $F(2, 244) = 0.42$, $p > .05$. In brief, there were no differences on the mu parameters of participants' response distributions across network setups.

To test whether the var parameter of participants' response distributions can be predicted by the different network setups (i.e., independent/shared/sequential), we run a linear mixed-effects model, with network setups as fixed effects. The random effect consisted of an intercept for each subject: $(1|subject)$. The overall model fit was not significant, $R^2(R_c^2) = 0.65$, and the marginal exploratory power of the social network factor was $R^2(R_m^2) = 0.002$, $F(2, 244) = 1.14$, $p > .05$. In brief, there were no differences on the var parameters of participants' response distributions across network setups.

To test whether the Jensen-Shannon Divergence ($D_{\text{JS}}$) of participants' response distributions can be predicted by the different network setups (i.e., independent, shared, sequential), we run another linear mixed-effects model, with network setups as fixed effects. The random effect consisted of an intercept for each subject: $(1|subject)$. The overall model fit was not significant, $R^2(R_c^2) = 0.121$, and the marginal exploratory power of the social network factor was $R^2(R_m^2) = 0.008$, $F(2, 244) = 0.18$, $p > .05$. In brief, there were no differences on the $D_{\text{JS}}$ of participants' response distributions across network setups.

### Quantitative comparison - Predictions vs Observed

Taken together, the results of the quantitative comparisons suggest that the model predictions and participants' responses were not systematically similar.

In order to check whether the mu parameter of participants' posterior distribution was systematically different than the mu of model posteriors, we run an unpaired t-test between participants' posterior mu's and model posterior mu's. The results revealed a significant difference between the mu parameters, $t(719.6997) = 3.656417$, $p < 0.001$, with participants mean mu being 0.56, while model prediction 0.51 (confidence interval of the difference $[0.02, 0.08]$). That is, the difference between participants' mu parameters and model mu parameters is not close to zero, with participants' judgements being closer to ethical than what model predicted.

To check whether the var parameter of participants' posterior distribution was systematically different than the var of model posteriors, we run a Wilcoxon signed-rank test (non-parametric t-test) between participants' posterior var's and model posterior var's. The results showed that there was a significant difference between the var parameters, $v = 120778$, $pp < 0.001$ with participants mean var being 0.014, while model prediction

0.002 (note here this is the *inverse* variance, meaning that smaller values indicate more confident responses. That is, the difference between participants' var parameters and model var parameters is not close to zero, with model prediction being more confident than participants'.

### 6.5.3 Quantitative comparison - Participants priors vs Participants posteriors

Overall, the results of the comparisons below seem to suggest that participants' prior judgements were not different than their posterior judgements. However, participants' certainty increased after being exposed to the experimental condition.

In order to evaluate whether there was a significant difference between the mu parameters of participants' prior and posterior distributions, we run an paired t-test between participants' prior and posterior mu parameters. The results showed that there was no significant difference between the mu parameters, $t(368) = -0.48$, $p = 0.633$, with the mean difference being very close to zero, (confidence interval of the difference $[-0.0266, 0.0162]$). That is, the difference between prior and posterior mu parameters is insignificant (fig. 6.8).

In order to check whether the var parameter of participants' posterior distribution was systematically different than the var parameter of their prior, we run a Wilcoxon signed-rank test (non-parametric t-test) between participants' posterior and prior vars. The results showed that there was a significant difference between the var parameters, $v = 32135$, $p < .001$ with participants prior mean var being 0.019, while posterior mean var 0.0136 (note here this is the *inverse* variance, meaning that smaller values indicate more confident responses. That is, the difference between participants' prior and posterior var parameters is not close to zero, with participants being more confident after being exposed to the experimental condition (fig. 6.9).

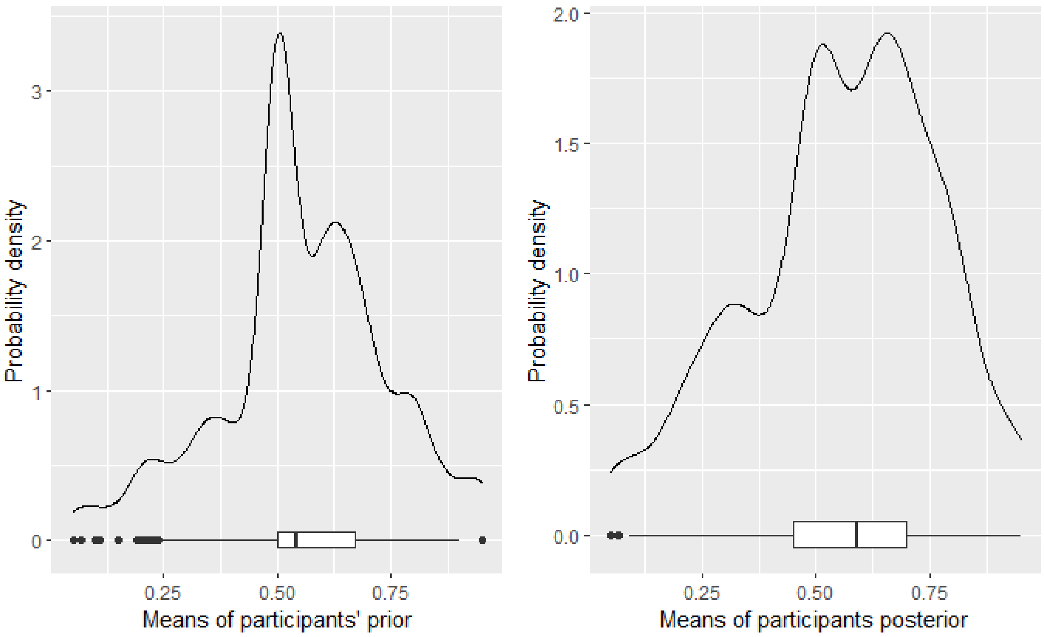


*Figure 6.8.* Probability density distributions for participants' prior and posterior mean (belief).

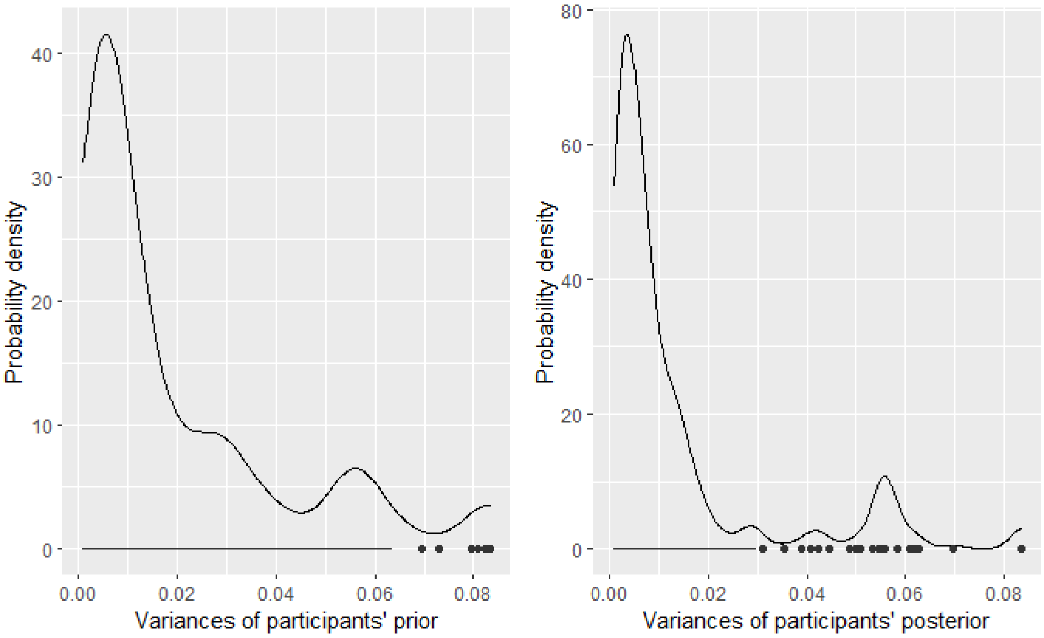


*Figure 6.9.* Probability density distributions for participants' prior and posterior variance (certainty).

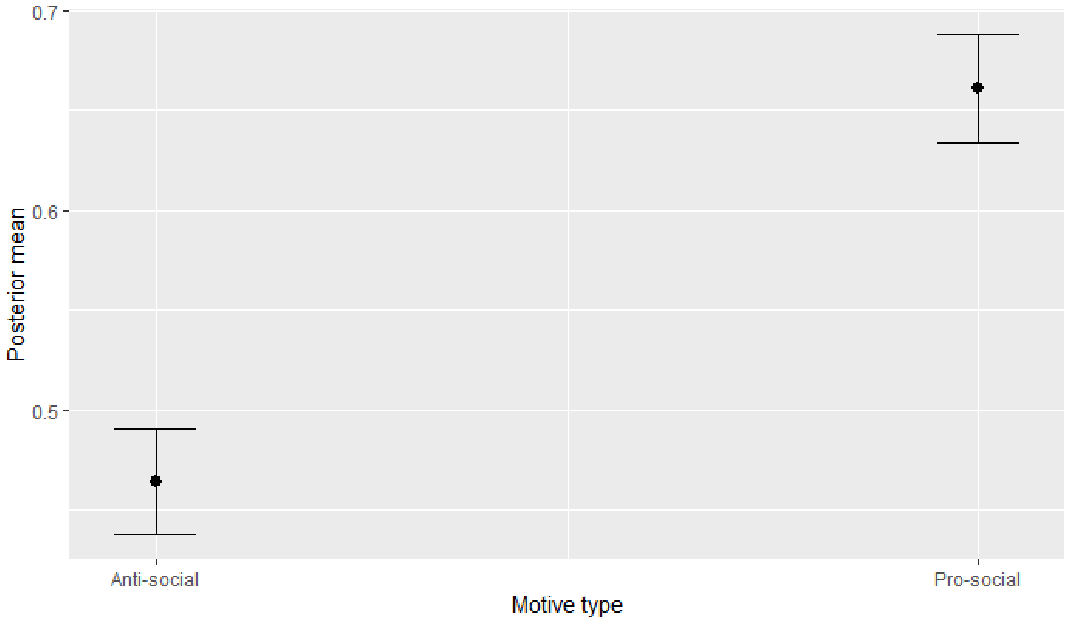


*Figure 6.10.* Participants' mean posterior for the mean parameter by motive type. Mean range 0 - 1, where 0 is the unethical side, while 1 is the ethical one. Pro-social scenarios trigger responses closes to the *ethical* end.

### 6.5.4 Quantitative comparison - Participants posteriors over motive type

To test whether the mu parameter of participants' response distributions can be predicted by the different motive types (i.e., anti-social/pro-social), we run a linear mixed-effects model, with response type as fixed effect. The random effect consisted of an intercept for each subject: (1—subject). It was found that the motive type did improve model fit significantly F(1, 367) = 104.48, $p$ < .001, with pro-social motives eliciting judgements closer to the ethical end (fig. 6.10).

To test whether the var parameter of participants' response distributions can be predicted by the different motive type (i.e., anti-social/pro-social), we run a linear mixed-effects model, with motive type as fixed effect. The random effect consisted of an intercept for each subject: (1—subject). The results showed that the amount of the parameter variance explained by the fixed effects is 0.27%, marginal r2 (r2m) = 0.0027, while the variance explained by the entire model is 65.07%, conditional r2 (r2c) = 0.65. It was found that the motive type did not improve model fit significantly F(1, 272.5769) = 2.18,

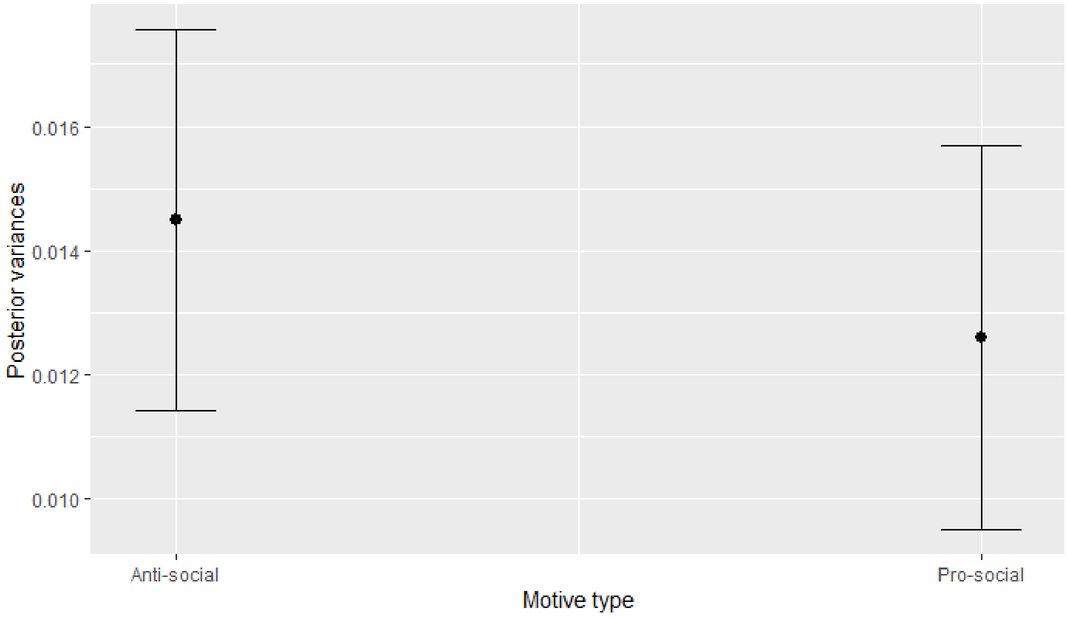


*Figure 6.11.* Participants' mean posterior for the var parameter by motive type. Mean range 0 - 1, where 0 is the unethical side, while 1 is the ethical one. There was no difference.

$p$ = 0.14, with no comparison across the mean var's being significant, all p's ¿.05. In brief, there were no differences on the var parameters of participants' response distributions across motive type (fig. 6.11).

### 6.5.5 Quantitative comparison - Participants belief revision in crossover condition

In order to test if participants' belief revision was significantly different than model prediction in the crossover condition, we run a Wilcoxon signed-rank test (non-parametric t-test) between participants' $D_{\mathrm{JS}}$ (difference between prior and posterior) and model $D_{\mathrm{JS}}$ (difference between prior and posterior) focusing only in the crossover subset (where participant's initial judgement was closer to one end (e.g., ethical), while participant's final judgement was closer to the other end (e.g., unethical)). The results showed that there was a significant difference between the $D_{\mathrm{JS}}$ parameters, $v = 3819$, $p < .001$ with participants mean var being 8.85, while model prediction 24.05. That is, in the crossover condition, participants' posterior responses were closer to their prior responses compared

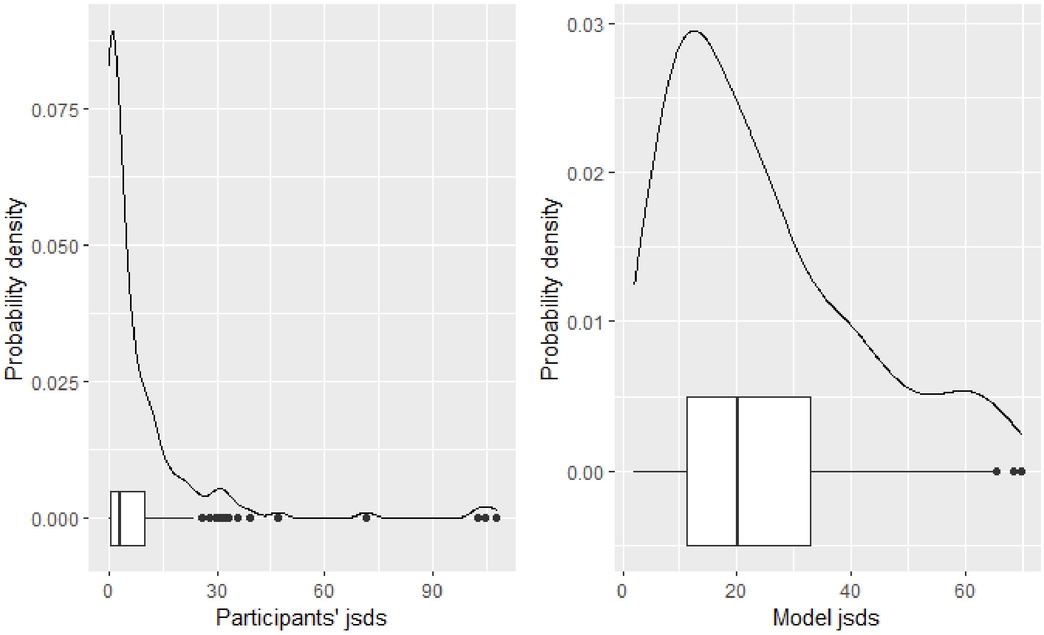


*Figure 6.12.* $D_{\text{JS}}$ probability densities of model predictions and participants' responses. Participants' $D_{\text{JS}}$ distribution surrounds 1-5, while model $D_{\text{JS}}$ surrounds 16-20. That is, participants' difference between prior and posterior responses is smaller than this of model predictions.

to what the model predicted, that is, participants seem to stuck with their prior (fig. 6.12).

## 6.6 Discussion

In this chapter, we modelled a moral belief updating process with two virtual social network neighbours across three different network setups. The three setups induced potential (in)dependencies that might exist between sources of social information. We found that subjects did not change their moral judgements, which is in line with previous work showing that moral judgements are resistant to revision in light of new evidence (e.g., Lord et al., 1979; Skitka, 2010). Here, we expanded these findings to the case in which new evidence comes in form of the judgements (experiment 1)/knowledge (experiment2) of social peers. Specifically, our Bayesian model —which predicted how an agent should update their judgements in an optimal way— failed to characterise participants' responses to information from peers. The model predicted a smaller difference between prior and posterior judgements when the judgements of social network neighbours were dependent as compared to the independent condition. Moreover, it predicted that the dependent case of the sequentially updated judgements would result in a smaller update of prior judgements as compared to shared information. However, in the first experiment, as expected from the subjectivity ingrained within moral judgements and their resistance to change, the empirical data did not support these normative predictions: participants did not update their moral judgements in any condition. This was true across all three metric comparisons (see Fig. 6.6, left column). Furthermore, posterior distributions could not be predicted by the three different network setups. Participants' posterior distributions were significantly different from model posteriors (see Fig. 6.6, right column). Contrary to our expectation, participants mu parameters followed a normative pattern, aligned to our normative model predictions.

The second experiment, on the other hand, showed a slightly different picture. When the response type (compatible/crossover) and/or motive type (anti-social/pro-social) are not taken into account then participants do not update their judgements after they are exposed to the experimental condition. However, when we factor in the effect of the response type, participants do update their judgements in the crossover condition, although they do so to a smaller extent than a rational agent (i.e., model prediction). In other words, participants seem to be resistant to altering their judgements even when subsequent evidence indicates their previous judgement/belief was erroneous (see also figure 6.12. As expected, participants' judgements were closer to *unethical* when the motive type of the main character of the scenario was anti-social and when the motive type was pro-social, participants' judgements were closer to *ethical.* Interestingly, the difference from the middle point of the unethical-ethical spectrum (neutral) was smaller for the anti-social motive type than the pro-social one. That is, participants characterised anti-social motives as unethical, but they did so less than they characterised pro-social motives as ethical. This aligns with pre-existing literature that supports that when people revise or form their beliefs are more cautious when judging others negatively (Ashton & Ashton, 1990).

Our findings are directly comparable to Fränken et al. (2020) where we used a slightly different paradigm (i.e., non-moral context) but the same framework and similar analyses. There, people update their judgements significantly less when the provided social information was coming from dependent sources (i.e., shared and sequential network setups) as compared to the independent case. That is, people were not simply combining their own judgements with the communicated beliefs of their network neighbours. Rather, they were additionally sensitive to the origin of those judgements and to what extent they were redundant. Here, by contrast, people seemed to ignore the origin and details of social information, simply sticking to their prior beliefs.

We thus replicated previous findings showing that moral judgements are most often strongly held convictions rooted in emotion and socialisation that are not changed easily by deliberative reasoning. There is, however, some support / nuance of optimism coming from Horne et al. (2015) (see also Schuldt et al., 2017), who found that, given certain types of evidence, people do revise their beliefs. A key feature of this evidence is that it must come ”from the inside” in order to induce a belief change. If people actively considered a moral dilemma while making a judgement, and they were presented with another (contradictory) dilemma and the chance to revise their judgements, people might adjust their belief. For example, Horne et al. (2013) asked participants to make a judgement about a moral dilemma and then immediately after to rate their agreement with different moral beliefs. It was found that considering the Footbridge dilemma –a traditional dilemma that tends to elicit deontological judgements– let people to lower their credence in the utilitarian belief: ”in the context of life or death situations, always take whatever means necessary to save the most lives” (Horne et al., 2013). They concluded that people revise their moral beliefs after exposure to a single dilemma.

A limitation of the present study is our assumption that the initial data (i.e., moral scenarios) comes from a binomial distribution. Although it allows for a tractable computation of the posterior, this assumption might not hold in reality. Further work might assess the impact of switching to different distributions (e.g., Gaussian) or try to infer the real distribution of the data in order to make the model more reflective of everyday life. Another extension of the current study would be to use friends or people from the real network of the subject, instead of virtual agents to increase the strength of the social evidence.

In summary, in the current study we used a novel method to derive the distributions of participants' moral judgements by asking them to provide mean and precision using a dynamic interface. Participants' responses did not align with our normative Bayesian framework predictions, replicated the finding that moral judgements are resistant change in light of new evidence, even if this evidence takes the form of social information from network peers.

# Chapter 7

# General Discussion

## 7.1 Our studies

The inherent challenge, and perhaps exciting promise, of political psychology and cognitive science is the task of exploring an extremely complex system –the cognitive agent– in an endlessly intricate space –the political space. These complexities inherently blend one another, rendering a powerful psychological science of ideology and political behaviours both challenging and vital. The speedy expansion of misinformation propagated by digital media as well as noticeable polarisation within and across national bodies has triggered a worldwide feeling that our knowledge about the origin of voting behaviours and ideological world views is perilously inadequate. Whereas the research of political beliefs and behaviours has been historically restricted to the social sciences, recent progress in computational cognitive sciences highlights that a computational approach might critically improve our understanding about political and ideological behaviours. Computational perspectives on the nature of ideological behaviours normally take two main forms: computational simulations of hypothetical behavioural dynamics and computational modelling of behaviour on cognitive tasks. The sections in this dissertation reflect both types of computational approaches and results in striking overlaps and complementary findings.

We first tried to capture how, in a political space, conflicting beliefs might lead to cognitive dissonance, and which strategies (e.g., disengagement) are being used by agents to overcome the dissonance they feel. Then, we investigated how conflicting environments might shape political polarisation, and finally, how people update their beliefs and departure from normative updating (where presumably conflict/dissonance might play a role).

## 7.2 Modelling the frequency of conflict using PGM

In chapter 2 we defined a probabilistic graphical model to investigate the frequency with which different political ideologies experience conflicting beliefs, and thus potentially disengage from some of these beliefs as a means to resolve the conflict. The main findings indicated that more conservative political ideologies lead to a higher probability of experiencing conflicting beliefs, and thus disengaging from those beliefs.

Looking closer, however, we saw that extreme conservatives' probability of experiencing conflict was slightly less than that of conservatives. This somewhat unexpected finding can be understood probabilistically if we look closer in the mathematical function we used to describe conflict. Based on that function, for a conflict to come about we need at least two *highly* and *equivalently* upheld beliefs. In other words, the difference between the two (or more) conflicting foundations works as a weighting parameter for the endorsement of the foundations. The more the difference, the less the endorsement. Applying that to conservatives (see figure 1.1), we see that the *difference* between any pair of the foundations should not be high since all foundations are *similarly* valued. On the other hand, extreme conservatives value the binding foundations (loyalty, authority, sanctity) more than the individualising ones (care, fairness). Therefore, the difference between the foundations should be greater, thus decreasing the probability of a conflict (between binding and individualising foundations).

Another important finding was about the frequency of conflict within different moral scenarios. Specifically, within individualising scenarios have the smallest probability to accommodate conflict, while between individualising and binding scenarios have the highest. Within binding foundations scenarios are intermediate in probability of eliciting conflict. As we explain in the following paragraph, this finding has important theoretical implications for research on echo chambers and ideological polarisation.

Although, for the most, we discuss conflict occurring *within* the agent, meaning, for example, that an agent could have contradictory beliefs, our conflict model is not restricted only to such scenarios. Theoretically, our model predictions compute potential conflict between foundations, be it within or between agents. Using that information alongside with event's moral foundations (i.e. the salient moral aspects of a given situation), the model can predict which moral scenario (e.g., within individualising foundation, within binding foundations, between individualising and binding foundations) is more probable to generate a conflict –either within or between agents. Findings like these could have implications for echo chambers, where people surround themselves with like-minded others who have the same or relevant beliefs, in an attempt to avoid conflict, and thereby reflect and reinforce their own beliefs (e.g., Garrett, 2009; Farrell, 2015; Petersen et al., 2013; Knobloch-Westerwick et al., 2020). We use the relevant findings as initial parameters for the agent-based model presented in chapter 5, where we investigate echo chamber formation between political ideologies.

## 7.3 Modelling the frequency of conflict – revision

In chapter 3 we revise the conflict model and update two of its aspects. Specifically, we 1) add interaction between the agent's *moral foundations*, and we 2) add the concept of agreement between *event's moral foundations*. In particular, in this version the activation of one foundation can influence the activation of another in a way that the two individualising foundations have a positive correlation between them, binding foundations are also positively correlated, while between individualising and binding foundations there is a negative relationship. Furthermore, in the revised version there is a higher chance of agreement between foundations coming from the same cluster rather than between clusters. This also in a sense allows for co-variation to exist but this time between the moral foundations of the event.

To allow for these changes we introduce a slightly different type of PGM, one that can accommodate connections between nodes, which cannot ascribe a directionality to the interaction between variables. This time we are using a *partially directed acyclic graph* (PDAG) which is capable of hosting both types of edges.

Having implemented these changes, the aim of the study remained the same: to explore how moral values might lead to conflicting beliefs, which in turn drive one to disengage from these beliefs. The findings generally replicated the ones from chapter 2. In general lines, the differences between the two versions of the model have only to do with the magnitude of the conflict and not with its pattern. In particular, the further to the right one scores on the political spectrum, the more likely one is to experience conflict. This pattern breaks, again, in extreme conservatives. This time, however, the steep drop in the probability of conflict from conservatives to extreme conservatives (seen in the previous chapter) became a smooth decrease. Regarding the conflict triggered within different conflicting scenarios, the findings again showed only some minor changes on the magnitude of the conflict. When only the event's moral foundations are observed, within individualising foundations have the smallest chance to generate conflict, between individualising and binding foundations scenarios have the highest, with within binding foundations scenarios being intermediate. All the other findings were relatively unchanged, with scarce differences on the magnitude of the conflict.

## 7.4 Moral disengagement – Empirical data

We then went on to validate the model and tested its predictions by collecting empirical data collected independently in Edinburgh, UK, and Napoli, Italy. In particular, in chapter 4, we tested whether the predictions presented by the conflict model hold true also in empirical data. We used the conflict model (chapters 2 and 3) to draw predictions. Besides testing the frequency of moral disengagement based on the political orientation of the agent, in chapter 4 we also add the right-wing authoritarianism (RWA) and social dominance orientation (SDO) scales in the equation, to see whether the relationship between political orientation and moral disengagement holds true even after controlling for these two variables.

Largely, the predictions of the conflict model were valid. The more to the right one scored on the political spectrum, the higher the tendency to self-report using disengagement mechanisms as a means to justify morally questionable actions. The fade-off pattern predicted by the conflict model from conservatives to extreme conservatives was only partially supported by the empirical data. In particular, to check whether the fade-off pattern emerged also in the empirical data, we created two regression models, one reflective and one non-reflective of the fade-off pattern. We then used two comparison criteria, AIC and BIC, which resulted into a tie. However, we argue that the discrepancy between the criterion might be an artefact of the weak power of our study, and that increasing the sample size could have captured the subtle effect better.

When right-wing authoritarianism and social dominance orientation effects are taken into account, political orientation seemed to lose predictive capacity. That is, the relationship between political orientation and moral disengagement is non-significant when also accounting for SDO and RWA. We argued that this could be because SDO and RWA are closely related to political orientation. In particular, RWA has its intellectual roots in issues of fascism (Adorno et al., 1950) and is characterised by submission to perceived authorities, aggression when sanctioned by authorities, and adherence to social conventions (B. Altemeyer, 1994; R. A. Altemeyer & Altemeyer, 1996; B. Altemeyer, 1998). SDO is rooted in Social Dominance Theory (Sidanius & Pratto, 2001; Pratto et al., 2006) which argues that social groups are hierarchically arrayed in societies in terms of status and resources, and conflict among groups is minimised, in part, through socially shared beliefs that justify the hierarchy (e.g., legitimising myths). SDO, in particular, is an anti-egalitarian orientation characterised by a desire for hierarchical relations among groups (Pratto et al., 1994; Sidanius et al., 1994).There are elements of conceptual overlap here not only with political ideology in the general sense, but also moral disengagement. Thus, overlapping predictive variance is not surprising.

All in all, the patterns from the empirical data support the conflict model predictions, with the possible exemption being the fade-off pattern, which was only partly supported.

## 7.5 Ideological polarisation

In chapter 5, we used the predictions of the conflict model as inputs to a simple agent-based model to investigate echo chambers and polarisation in different political ideologies. We focused on just two (of many possible) factors responsible for echo chamber formation: 1) confirmation bias (i.e., the need to confirm one's beliefs, frequently by associating with like-minded others), and 2) cognitive dissonance generated via interactions *between* agents and the need to decrease it. These two factors have opposing effects on the agent. In particular, 1) attracts them towards similar others, and 2) pushes them away from similar others (who made different choices). Using our conflict model predictions, we can differentiate between liberals and conservatives on the grounds of how frequently they clash with similar others. In our simulations, we found conservatives had a higher chance than liberals of clashing with ideological near-neighbours. More frequently conflicting agents should lead to a less dense sub-population than agents who do not clash with each other as frequently, that is, conservative groups should be less tightly bounded than liberal groups.

The results showed that the sub-population structures are controlled, in terms of group sizes and stability, by the density of the agents. The feedback loop between mobility and conflict yields a strong assortativity between physical and ideology space: closer neighbours tend to share the same ideology; this scenario is supported by the theory of confirmation bias. Furthermore, we saw that echo chambers emerge by individuals who can be influenced only by peers sharing similar views, although in echo chambers polarised opinions can co-exist. Last, the results showed that a higher chance of clashing with like-minded others leads to more restless agents, and thus, less dense sub-groups.

The contributions of the model are threefold. First, it shows that spatial segregation can result from dynamics involving agents seeking consensus on ideology. Second, it provides a framework in which the sub-group structure, often assumed in the modelling of social systems, emerges from the microscopic rules of the model itself. Third, it shows that conflicting agents yield the possibility that different opinions coexist within the same sub-group. Furthermore, the current study is the first (to our knowledge) that explores echo chamber formation looking at the need people have to avoid people with whom they share similar beliefs, rather than people with whom do not share similar beliefs, which is extensively looked at by other studies (e.g., Del Vicario et al., 2016; Ngampruetikorn & Stephens, 2016; Starnini et al., 2016).

## 7.6 Belief revision in a micro-social network

In the last chapter, chapter 6, we explored how individuals revise their moral beliefs in light of new evidence. We use a Bayesian model to detail normative predictions of a rational agent, and then we deployed two experiments to test human approximation to, or deviation from, that normative standard. In particular, we used three different social network setups, underlying three potential environments where an individual might update their beliefs based on what their social neighbours believe. In the first condition, 1) the agent communicates with two neighbours independently, and the two agents have formed their beliefs based on independent sources. In the second condition, 2) the two neighbours have shaped their beliefs based on the same source, while in the last condition, 3) although the two neighbours have formed their beliefs based on independent sources, they talk to each other prior to sharing their beliefs to the agent (see also figure 6.1).

The normative model acts as a *rational* agent in our study. That is, its predictions indicate the optimal way an individual should update their beliefs based on new evidence. The results align with previous work (e.g., Lord et al., 1979; Skitka, 2010; Fränken & Pilditch, 2020), that is, we found that people deviated from the rational norm, when updating their moral beliefs in light of new evidence.

More specifically we found that subjects did not change their moral beliefs at all. This expands pre-existing findings to the case in which new evidence comes in form of the beliefs of social network peers. Specifically, our Bayesian model —which predicted how an agent should update their beliefs in an optimal way— failed to characterise participants' responses to information from peers. The difference between prior and posterior beliefs was expected to be smaller when the beliefs of social network neighbours were dependent as compared to the independent condition. Moreover, the dependent case of the sequentially updated beliefs should have resulted in a smaller update of prior beliefs as compared to shared information. However, as expected from the subjectivity ingrained within moral beliefs and their resistance to change (e.g., Lord et al., 1979; Skitka, 2010), the empirical data did not support these normative predictions: in neither condition did participants update their moral beliefs. This was supported by all three metric comparisons (see figure 6.4,left column).

## 7.7 Implications

This work contributes to an important and interrelated set of questions on a number of levels of analyses. One of the main contributions is the demonstration of how approaches from social psychology could assist extend the range of formalisation in political sciences. We created a theoretical model of how frequently people adjust their political and moral beliefs to minimise cognitive dissonance — the distress that emerges when beliefs clash with pre-existing preferences. Having its roots in social sciences, this plain intuition describes why individuals regularly alter their choices to bring them into closer alignment with their actions/beliefs, or, when they do not control the environment that sparks the cognitive dissonance they feel, how they migrate from it to avoid experiencing such conflict.

Besides theoretical implications, these results might also help bridge the gap between liberals and conservatives both in a political level and in the day-to-day life. Politicians could be advised by our results by looking how frequent political ideologies conflict with each other and which foundations are responsible for these conflicts. This could help them prepare speeches, and therefore acts, that instead of fuelling the chasm between their voters they could decrease it by implementing more unifying solutions that are less conflicting. Another issue that our results might be able to help resolve is the online "fights" observed between individuals under articles posted on social media for a variety of matters, including abortion, capital punishment, gay rights, women's rights, gun ownership, environmentalism, euthanasia, and many more. Usually, liberals and conservatives adopt different standpoints on such matters, and this –in combination with the provocative content of the average article– creates a fertile ground for conflicts. The authors of such articles could get advised by our results and adjust the content of their article in a way that it will elicit less dispute between different political ideologies.

We finish by demonstrating that our framework is amenable to interposing new notions from social psychology that are tightly relevant to cognitive dissonance, such as confirmation bias and motivated reasoning (Lodge & Taber, 2013). Confirmation bias takes place when an individual refuses to look for possibly conflicting evidence, choosing instead to revise their beliefs based on evidence that aligns with their pre-existing beliefs (Schelling, 1969, 1971; Lewicka, 1988, 1998; Klayman & Ha, 1987; Cosmides, 1989; Del Vicario et al., 2016; Ngampruetikorn & Stephens, 2016; Starnini et al., 2016). In the same vain, motivated reasoning happens when an individual explicitly regards new information as completely congruent with their pre-existing beliefs (Druckman & Bolsen, 2011; Taber & Lodge, 2006). Either of these concepts are, at their core, scenarios of when people seek to dodge cognitive dissonance. With confirmation bias, cognitive dissonance is minimised by evading possibly challenging evidence that can trigger mental distress; with motivated reasoning, objective evidence is selectively disregarded also to decrease such distress. Either may be regarded as a special case of the larger framework that we propose here, that is, our approach could be used to formalise these increasingly critical concepts and describe their consequences.

## 7.8 Future directions

There are numerous ways to build upon the core models, which could potentially occupy future work on the subject. One possible feature beyond the scope of the present work is the addition of the concept of time (and thus learning). We saw that a political agent will experience conflicting beliefs with a given probability (chapters 2 and 3), and that they will avoid conflict-triggering environments (chapter 5), but how does the agent learn not to visit this type of environment in the future (i.e., learn to identify conflict-provoking environmental cues) or to strategically change their beliefs to avoid conflict? Cognitive dissonance theory proposes that the greater the magnitude of dissonance one experiences, the stronger the need to reduce it, and if an agent chooses to alter their beliefs to resolve dissonance, the greater the duration that change should last (Festinger, 1962; Festinger & Carlsmith, 1959). Moreover, moral disengagement theory also regards time as playing a crucial role in the duration one disengages from their beliefs (Bandura, 2002; Bandura et al., 2001). Therefore, introducing the time concept in our models would render them more representative of the relevant theories. This has great potential value, as studying the evolution of belief change, or moral disengagement, in complex social groups is extremely resource intensive and difficult. Particularly in the latter case, legal obfuscation almost always prevents any sort of ex post facto examination of the trajectory of disengagement. By contrast, building temporal dynamics into models of disengagement and belief change could provide a test-bed for predictions of how these processes unfold in the wild.

Another extension is to explore other micro-dynamic emergent phenomena for social groups, using the framework developed in chapter 5. An interesting example is to look at how misinformation spreads across polarised groups, and what are the implications in terms of how the spread shapes the group dynamics. Democrats or liberals tend to believe that human activity is a primary cause of climate change, whereas Republicans or conservatives are much less likely to hold this belief (Druckman & McGrath, 2019). As we saw in chapter 6, a prominent explanation for this divide is that it stems from directional motivated reasoning: individuals reject new information that contradicts their standing beliefs. Recent research (A. Moore et al., 2021) has showed that belief in the information is significantly affected by both (dis)trust in information source and by belief compatibility with the valence of the information. Neuroimaging results also confirm this pattern supporting that motivated cognition accounts of misinformation endorsement. What effects does this have on group segregation? Does new information acceptance play a role in how people choose their surroundings? This work would potentially suggest a new research agenda on climate change preference formation, and could have implications for effective communication.

## 7.9 Closing

In the current dissertation we used computational techniques as a tool to explore sociocognitive phenomena relevant to political ideologies. The results of this thesis expand the current research in social and cognitive fields on many levels. The findings should be used for explaining human behaviour to advance social science theory, rather than predicting human behaviour for practical purposes.

# Appendix A

# Appendices

## A.1 Probability theory

## A.2 Probabilistic graphical models framework

Probabilistic graphical models (PGMs) use graph-based representations to define complex joint distributions over random variables which are parameterised by an undirected or directed graph where edges between variables indicate statistical dependencies. Formally, we are given a set $X = \{X_1, ..., X_n\}$ of random variables where each variable $X_i$ takes a value $x_i$ from the domain $\mathcal{X}$. The domain $\mathcal{X}$ can be discrete, binary or real-valued, making $X_i$ a categorical, Boolean or continuous variable. The vector $x =< x_i, ...x_n >$ symbolises the joint assignment where each $X_i = x_i$. The goal is to define a probability distribution $P(X)$ over the joint assignment to all variables $X$. A PGM is defined by a graph $G = (X, E)$ whose nodes –or, more generally, vertices– correspond to the variables $X$. An edge $e_i \in E$ is either of the form $X_i \rightarrow X_j$ (directed) or $X_i$—$X_j$ (undirected) and indicates that $P(X)$ should model probabilistic dependencies between $X_i$ and $X_j$. Therefore, an important property of probabilistic graphical models is that the graph structure includes *conditional independencies* among variables, allowing the joint distribution to be compact without requiring all $2^n$ dependencies. Probabilistic graphical models are characterised by undirected or directed acyclic graphs (DAG). When $G$ is undirected, the resultant graphical model is referred to as a *Markov random field* (MRF), while the $G$ is a directed acyclic graph, the resulting PGM is a *Bayesian network* (BN). Follows a detailed explanation of these two terms and their conditional independence semantics in detail (focusing more on the Bayesian Networks since the current study uses only that).

### A.2.1 Markov random fields

Markov random fields (MRF) models are useful in modelling a variety of phenomena where one cannot naturally assign a direction to the relationship among variables. These undirected models are different than directed ones (see appendix A.2.2 - Bayesian networks) in that they offer a different and many times simpler perspective, in terms of both the independence structure and the inference task. However, these are too

Given an undirected graph $G = (X, E)$, and the related depth of maximal cliques $C = c_i, ..., c_M$ made by the edges $e \in E$, an MRF establishes the joint distribution over $x$ as:

$$P(x) = \frac{1}{Z} \prod_{k=1}^{M} \Phi_k(X_k) \tag{A.1}$$

where $Xk = x_j | x_j \in C_k$ is the set of all variables that take part in the $k$-th clique, and the vector $x_k$ is the assignment to the $X_k$ variables. $Z$ is the *log partition function* –a normalisation constant that requires exponentially many sums to compute, as we can see:

$$Z = \sum_{x_i \in X} \prod_{i=1}^{K} \Phi_k(X_k) \tag{A.2}$$

Therefore, evaluating $Z$ is many times intractable, but several useful approximations exist. The final component of an MRF are the $\Phi_k$ functions:

$$\Phi_k(X_k) = \exp(\lambda_k^T f_k(X_k)) \tag{A.3}$$

These functions are also know as the *clique potentials.* They have the property that $\log(\Phi_k)$ is linear and $\Phi_k$ is defined by a vector of feature function $fk(X_k)$ and weight vector $\lambda_k$. Each function $f_k^i(X_k)$ ascribes a real-value in $(0, \infty)$ to $x_k$ that measures the compatibility of this assignment to the variables $X_k$. Intuitively, higher scoring assignment configurations are exponentially more probable under the distribution. The set of all weight vectors $A = \{\lambda_k\}_{k=1}^{M}$ are the parameters of the MRF.

The undirected graph $G$ which defines an MRF model entails a few independencies among the variables in it. To explore these independences, we consider the neighbours of variable $X_i$, $Nei(X_i)$ (i.e., the nodes connected to $X_i$ by an edge), in the $G$. The local Markov property of distribution $P$ with respect to $G$ indicates that each $X_i$ is conditionally independent of variables $X \backslash X_i$ given its neighbours $Nei(X_i)$, denoted $(X_i \perp X \setminus X_i \mid Nei(X_i))$. In undirected graphs, the neighbours $X_i$, $Nei(X_i)$, denotes its *Markov blanket*, that is, the set of variables required to render $X_i$ conditionally independent of other variables in the graph.

All in all, Markov networks, similar to Bayesian networks, may be regarded as specifying a group of independence assumptions indicated by the graph structure. Markov (i.e., undirected) graphs can be seen as a data structures for defining a probability distribution in a factorised form. As we saw from the equation A.1 the factorisation is given as a product of factors (generic non-negative functions) across cliques in the system. Regarding the independence assumptions introduced by the graph, there are various likely definitions that are similar for positive distributions. For directed graphs, there is a different factorisation of the joint distribution and as a consequence, there are different kind of conditional independences as we see in the next section.

### A.2.2 Bayesian Networks

Bayesian networks (BNs) build upon equivalent ideas as Markov random fields models by utilise conditional independence qualities of the distribution to enable concise and natural representations. Despite that, the obey different types of conditional independences since the edges have a direction. They provide us the flexibility to adapt our depiction of the distribution to the independence qualities that seem sensible in the given context.

The cornerstone of a BN is a *directed acyclic graph* (DAG) $\mathcal{G}$, with units being random variables of the system and with edges corresponding to direct effects of one node on others. The graph $\mathcal{G}$ can be seen in two distinct ways:

- as a data structure that offers the foundation for depicting a joint distribution concisely in a factor-friendly manner.
- as a concise depictions for a group of conditional independence assumptions regarding a distribution.

These two views are strongly equivalent.

Let us now consider the joint distribution of a BN. Given a DAG $\mathcal{G} = (\mathcal{X}, \mathcal{E}$ and a function $\pi(X_i)$ that maps $X_i$ to its parerents in $\mathcal{G}$ (variables with edges incoming to $X_i$), a BN defines the joint distribution over $X$ as:

$$P(x) = \prod_{i=1}^{n} p(x_i \mid \pi(X_i)) \tag{A.4}$$

where the conditional probabilities $p(x_i \mid \pi(X_i))$ parameterise the distribution. When variables $X$ are categorical or Boolean, these conditional probabilities can be represented by conditional probability tables (CPDs). The BN defined by $\mathcal{G}$ fulfils two conditional independence properties: the local and the global Markov property. The global Markov property of a distribution relies on all the other conditional independences entailed by $\mathcal{G}$. The local Markov property of a distribution is that each $X_i \in X$ is independent of its non-descendants in $\mathcal{G}$ conditioned on its parents $\pi(X_i)$. The independence entailment criteria on the graph is known as d-seperation and builds on the notion of blocked paths.

### A.2.3 Partially directed models

Previously, we described two differing kinds of graphical models, on the basis of directed and undirected graphs. We can merge them by enabling systems to have either directed and/or undirected relationships. Generally, there are two types of partially directed models, the *partially directed acyclic graphs* (PDAG) or *chain graphs*, and the *conditional random field* (CRF) models. PDAG is a generalisation of the CRF models, which introduce a whole model where undirected elements rely on one another in a directed manner. In the current we use and describe only PDAG.

Partially directed acyclic graphs are employed to offer a generic solution, –or better yet, handling– of the independence assumptions induced in PDAGs. In PDAGs the units can be dis-jointly divided into multiple *chain components*. Links among units *within* the same chain component have to be undirected, whereas links among nodes in *different* chain components must be directed. Therefore, PDAGs are also named chain graphs.

Naturally, every chain component $K_i$ in the graph is related with $P(K_i \mid Pa_{K_i})$ – the conditional distribution of $K_i$ given its parents in the graph. In particular, every node is specified with a list of factors that include the variables in $K_i$ and their parents; the distribution $P(K_i \mid Pa_{K_i})$ is defined using the factors related with $K_i$ to specify a CRF with target variables $K_i$ and observable variables $Pa_{K_i}$. To give a formal definition, it might help to briefly describe the notion of a *moralised* directed graph.

If $\mathcal{G}$ is a PDAG and $G_1, ..., G_l$ are its chain elements; we specify $PaG_i$ to be the parents of the units in $G_i$. The moralised graph of $\mathcal{G}$ is an undirected graph $\mathcal{M}[\mathcal{G}]$ made by 1) linking –only with undirected edges– every pair of units $X, Y \in PaG_i$ for every $i = 1, ..., l$, and then transforming every directed edge into an undirected edge.

This description generalises the notion of a moralised directed graph. For directed graphs, every unit is its own chain element, and thus we are just adding undirected links among the parents of every node.

Last, we need to define the factorisation of a chain graph. If $\mathcal{G}$ is a PDAG, and $G_1, ..., G_l$ are its chain elements. A chain graph distribution is specified over a list of factors $\phi_i(D_i)(i = 1, ..., m)$, in a way that every $D_i$ is a whole sub-graph in the moralised graph $\mathcal{M}[\mathcal{G}^+[D_i]]$. We relate every factor $\phi_i(D_i)$ with a sole chain element $G_j$, in a way that $D_i \subseteq G_i \cup PaG_i$ and defines $P(G_i \mid PaG_i)$ as a CRF with these factors, and with $Y_i = G_i$ and $X_i = PaG_i$. We can now define

$$P(\mathcal{X}) = \prod_{i=1}^{l} P(G_i \mid PaG_i).$$

Then a distribution $P$ factorises across $\mathcal{G}$ if it can be represented as a chain graph distribution across $\mathcal{G}$.

### A.2.4 Inference

#### Conditional probability query

We are now turning our attention on the problem of performing inference in graphical models. One of the most usual form of query is the *conditional probability* query, $P(Z \mid Y = y)$, which is used for various helpful reasoning patterns, such as explanations, predictions, inter-causal reasoning, to name a few examples. Formally, we define conditional probability as:

$$P(Z \mid Y = y) = \frac{P(Z, y)}{P(y)} \tag{A.5}$$

Every instantiation of the numerator is a probability manifestation P(z, y) that can be calculated by marginalising out all inputs in the joint distribution that match with assignments compatible with $z, y$. In particular, if $W = \mathcal{X} - Z - Y$ are the random variables that are not either query or evidence, then we have:

$$P(z, y) = \sum_{w} P(z, y, w) \tag{A.6}$$

Since $Z, Y, W$ are all the variables of the system, every expression $P(z, y, w)$ in the summation is just an input in the joint.

The probability $P(y)$ can be calculated with two ways: either directly by summing out the joint, or, more conveniently:

$$P(y) = \sum_{y} P(z, y) \tag{A.7}$$

which allows us to cache and re-use the results of equation A.6. If we calculate both equation A.6 and equation A.7, we can then divide each $P(z, y)$ by $P(y)$, to obtain the wanted conditional probability $P(z \mid y)$. Notice that this procedure matches to getting the vector of marginal probabilities $P(z^1, y), ..., P(z^k, y)$, where $k = |Val(Z)|$ and *renormalising* the inputs to sum to 1.

#### Maximum a posteriori

Another quite used inference problem is the *Maximum a posteriori* (MAP) which targets to get the most probable assignment to all of the non evidence variables. Additionally, a *marginal* MAP query tries to get the most probable assignment to a sub-set of the variables, marginalising out across the rest.

Maximum a posteriori queries are usually used as a method of *filling in* unknown information. Take for instance a message decoder which tries to decode messages transmitted over a noisy channel. After observing a sequence of noisy bits, the decoder tries to get the most probable assignment of entered bits that may have produced the given observation. To tackle this type of inference problem, it is better to use MAP query instead of a standard probability one, as we are not interested in the most likely values for the individual bits sent, but rather in the message whose overall probability is highest.

Formally, MAP inference corresponds to the optimisation:

$$x^* = \arg\max_x P(x) \tag{A.8}$$

It is standard to maximise $\log P(X)$, which gives an equivalent solution. In MRFs, the normalisation constant can be ignored and $\log \prod_{k=1}^{M} \Phi_k(X_k)$ is instead maximised. Generally speaking, MAP inference in graphical models is NP-hard (i.e., the method for solving this problem can be translated into one method for solving any NP-problems, thus, it s "at least as hard as any NP-problem").

### A.2.5 Learning

Our discussion so far assumes that we have a given graphical model. Given a graphical model, we then make inferences or we reason about different aspects of the underlying phenomenon by using conditional independencies, assuming that the model –structure and parameters– was part of the input. Of course that is not necessarily the case.

There are two approaches to the task of having a model. The first is to construct the network by hand, which is also the approach we have adopted in this project. The construction of a network, however, requires a skilled knowledge theoretician who spends several days (or in my case, a PhD worth of time) with one or more domain experts (e.g., supervisors).

Such manual network development could be problematic in several cases. In some cases, the volume of knowledge needed is simply too high or the experts' time is too important. In other cases, there is just no expert who has adequate grasp of the area. In various areas, the qualities of the distribution alter from one application site to another or over time, and it is unfeasible for an expert to re-design the system every month. Therefore, what we need is access to a sample of instances from the distribution we want to model, which we let our model figure our the parameters, or even its structure. This process of building a model up from a list of instances is broadly called *model learning*.

The choice of graphical model affects the complexity of the learning problem. Since for MRFs the log partition function $Z$ depends on the parameters, it cannot be omitted from the optimisation. In general, learning the parameters of arbitrarily cyclic MRFs requires approximations to the full log likelihood. In BNs, one the other hand, the log likelihood decomposes into a sum over the conditional probability for each variable $X_i$, admitting exact and often closed-form solutions.

Another crucial learning problem in PGMs is the structure learning which tries to learn the underlying graph $\mathcal{G}$ from observed data. Structure learning corresponds to model discovery and plays a critical role in computational science, but we are not going to cover it in this project. The interested reader can refer to resources such as Koller and Friedman (2009), Zhou (2011) and Gasse (2017) and more.

## A.3 Conditional probability distribution over PO and agent's MF

Table A.1
Conditional probability distribution over political orientation and agent's moral foundation – individualising foundations.

| | | **PO** | | | | |
|---|---|---|---|---|---|---|
| | | v. liberal | liberal | neutral | conservative | v. conservative |
| **MF** | disabled | 0.1 | 0.1 | 0.15 | 0.2 | 0.3 |
| | low | 0.2 | 0.2 | 0.3 | 0.4 | 0.4 |
| | medium | 0.3 | 0.4 | 0.35 | 0.3 | 0.2 |
| | high | 0.4 | 0.3 | 0.2 | 0.1 | 0.1 |

MF (moral foundations) conditioned on PO (political orientation) $P(MF \mid PO)$.

Table A.2
Conditional probability distribution over political orientation and agent's moral foundation – binding foundations.

| | | PO | | | | |
|---|---|---|---|---|---|---|
| | | v. liberal | liberal | neutral | conservative | v. conservative |
| **MF** | disabled | 0.4 | 0.4 | 0.25 | 0.1 | 0.1 |
| | low | 0.4 | 0.3 | 0.4 | 0.2 | 0.1 |
| | medium | 0.1 | 0.2 | 0.2 | 0.3 | 0.4 |
| | high | 0.1 | 0.1 | 0.15 | 0.4 | 0.4 |

MF (moral foundations) conditioned on PO (political orientation) $P(MF \mid PO)$

# A.4 Example scenarios

## A.4.1 Experiment 1

*Ann is the human resources manager for a large accounting firm. The firm is currently looking for a new certified public accountant (CPA), and Ann has already interviewed several applicants for the position. There are four applicants vying for the position, and one of them happens to be a member of a church that Ann goes, so they know each other from various church activities. However, one of the other applicants has slightly better credentials and a little more experience than Ann's fellow church member does. Ann want to hire the candidate who belongs to her church, as Ann knows she shares her values and beliefs and almost feels like "family," but Ann also feels that she might be cheating the other applicant out of a job that she was legitimately qualified for, and perhaps even more qualified for. Ann hires the applicant who is a member of her church.*

# Bibliography

Aaron, P. (2010). Parler is the new twitter for conservatives. here's what you need to know. Retrieved February 6, 2021, from https://fortune.com/2020/06/29/what-is-parler-app-social-media-conservatives-who-owns-free-echo-facebook-twitter-verified-faq/

About-Facebook. (2019). Why am i seeing this? we have an answer for you - about facebook. Retrieved February 1, 2021, from https://about.fb.com/news/2019/03/why-am-i-seeing-this/

Abramowitz, A. I. & Saunders, K. L. (2008). Is polarization a myth? *The Journal of Politics*, *70*(2), 542–555.

Adorno, T., Frenkel-Brunswik, E., Levinson, D. & Sanford, R. (1950). The authoritarian personality harper and row. *Publishers New York*.

Alam, S. J. & Geller, A. (2012). Networks in agent-based social simulation. *Agent-based models of geographical systems* (pp. 199–216). Springer.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of personality and social psychology*, *49*(6), 1621.

Altemeyer, B. (1994). Reducing prejudice in right-wing authoritarians. *e Psychology of Prejudice: e Ontario Symposium*, *7*, 131–148.

Altemeyer, B. (1998). The other "authoritarian personality". *Advances in experimental social psychology* (pp. 47–92). Elsevier.

Altemeyer, B. (2007). The authoritarians: University of manitoba.

Altemeyer, R. A. & Altemeyer, B. (1996). *The authoritarian specter*. Harvard University Press.

Altemeyer, R. A. & Altemeyer, B. (1981). *Right-wing authoritarianism*. University of Manitoba press.

Andrews, M., Vigliocco, G. & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review*, *116*(3), 463.

Ansolabehere, S., Rodden, J. & Snyder Jr, J. M. (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, 215–232.

Anunrojwong, J. & Sothanaphan, N. (2018). Naive bayesian learning in social networks. *Proceedings of the 2018 ACM Conference on Economics and Computation*, 619–636.

Aquino, K., Reed II, A., Thau, S. & Freeman, D. (2007). A grotesque and dark beauty: How moral identity and mechanisms of moral disengagement influence cognitive and emotional reactions to war. *Journal of Experimental Social Psychology*, *43*(3), 385–392.

Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A. & Batty, M. (2015). Constructing cities, deconstructing scaling laws. *Journal of the royal society interface*, *12*(102), 20140745.

Arcaute, E., Molinero, C., Hatna, E., Murcio, R., Vargas-Ruiz, C., Masucci, A. P. & Batty, M. (2016). Cities and regions in britain through hierarchical percolation. *Royal Society open science*, *3*(4), 150691.

Ashton, R. H. & Ashton, A. H. (1990). Evidence-responsiveness in professional judgment: Effects of positive versus negative evidence and presentation mode. *Organizational Behavior and Human Decision Processes*, *46*(1), 1–19.

Astrove, S. L., Yang, J., Kraimer, M. & Wayne, S. J. (2015). Psychological contract breach and counterproductive work behavior: A moderated mediation model. *Academy of Management Proceedings*, *2015*(1), 11094.

Auchincloss, A. H. & Garcia, L. M. T. (2015). Brief introductory guide to agent-based modeling and an illustration from urban health research. *Cadernos de saude publica*, *31*, 65–78.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. *Simulating social phenomena* (pp. 21–40). Springer.

Axtell, R. (2000). Why agents?: On the varied motivations for agent computing in the social sciences.

Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, *113*(42), 11823–11828.

Bailey, R. M., Carrella, E., Axtell, R., Burgess, M. G., Cabral, R. B., Drexler, M., Dorsett, C., Madsen, J. K., Merkl, A. & Saul, S. (2019). A computational approach to managing coupled human–environmental systems: The poseidon model of ocean fisheries. *Sustainability Science*, *14*(2), 259–275.

Baldassarri, D. & Bearman, P. (2007). Dynamics of political polarization. *American sociological review*, *72*(5), 784–811.

Bandura, A. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ*, *1986*, 23–28.

Bandura, A. (1990a). Mechanisms of moral disengagement in terrorism. *Origins of terrorism: Psychologies, ideologies, states of mind*, 161–191.

Bandura, A. (1990b). Selective activation and disengagement of moral control. *Journal of Social Issues*, *46*(1), 27–46.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and social psychology review*, *3*(3), 193–209.

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of moral education*, *31*(2), 101–119.

Bandura, A., Barbaranelli, C., Caprara, G. V. & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of personality and social psychology*, *71*(2), 364.

Bandura, A., Caprara, G. V., Barbaranelli, C., Pastorelli, C. & Regalia, C. (2001). Sociocognitive self-regulatory mechanisms governing transgressive behavior. *Journal of personality and social psychology*, *80*(1), 125.

Bandura, A., Kurtines, W. M. & Gewirtz, J. (1991). Handbook of moral behavior and development. *Handbook of Moral Behavior and Development*, *1*, 45–103.

Banerjee, A., Breza, E., Chandrasekhar, A. G. & Mobius, M. (2019). *Naive learning with uninformed agents* (tech. rep.). National Bureau of Economic Research.

Barkun, M. (2013). *A culture of conspiracy: Apocalyptic visions in contemporary america* (Vol. 15). Univ of California Press.

Baron, J. & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the united states equally biased? *Perspectives on Psychological Science*, *14*(2), 292–303.

Baron, R. A., Zhao, H. & Miao, Q. (2015). Personal motives, moral disengagement, and unethical decisions by entrepreneurs: Cognitive mechanisms on the “slippery slope”. *Journal of Business Ethics*, *128*(1), 107–118.

Barrett, C. L., Eubank, S. G. & Smith, J. P. (2005). If smallpox strikes portland... *Scientific American*, *292*(3), 54–61.

Barsky, A. (2011). Investigating the effects of moral disengagement and participation on unethical work behavior. *Journal of business ethics*, *104*(1), 59.

Bartels, L. M. (2000). Partisanship and voting behavior, 1952-1996. *American Journal of Political Science*, 35–50.

Bartlett, B. (2015). How fox news changed american media and political dynamics. *Available at SSRN 2604679*.

Bates, D. M. (2010). Lme4: Mixed-effects modeling with r.

Baumann, F., Lorenz-Spreen, P., Sokolov, I. M. & Starnini, M. (2020). Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, *124*(4), 048301.

Beam, M. A. & Kosicki, G. M. (2014). Personalized news portals: Filtering systems and increased news exposure. *Journalism & Mass Communication Quarterly*, *91*(1), 59–77.

Beaudoin, C. A., Cianci, A. M. & Tsakumis, G. T. (2015). The impact of cfos' incentives and earnings management ethics on their financial reporting decisions: The mediating role of moral disengagement. *Journal of business ethics*, *128*(3), 505–518.

Becker, J., Brackbill, D. & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, *114*(26), E5070–E5076.

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G. & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, *10*(2), e0118093.

Beu, D. S. & Buckley, M. R. (2004). This is war: How the politically astute achieve crimes of obedience through the use of moral disengagement. *The Leadership Quarterly*, *15*(4), 551–568.

Bikhchandani, S., Hirshleifer, D. & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, *100*(5), 992–1026.

Boardley, I. D. & Kavussanu, M. (2011). Moral disengagement in sport. *International Review of Sport and Exercise Psychology*, *4*(2), 93–108.

BOBBIO, A., NENCINI, A. & SARRICA, M. (2011). Running title: Versione italiana del mfq.

Boutyline, A. & Vaisey, S. (2017). Belief network analysis: A relational approach to understanding the structure of attitudes. *American journal of sociology*, *122*(5), 1371–1447.

Brand-Ballard, J. (2003). Consistency, common morality, and reflective equilibrium. *Kennedy Institute of Ethics Journal*, *13*(3), 231–258.

Brandt, M. J., Sibley, C. G. & Osborne, D. (2019). What is central to political belief system networks? *Personality and Social Psychology Bulletin*, *45*(9), 1352–1364.

Brehm, J. W. & Cohen, A. R. (1962). Explorations in cognitive dissonance.

Briggs, R. (2014). Normative theories of rational choice: Expected utility.

Bruch, E. & Atwell, J. (2015). Agent-based models in empirical social research. *Sociological methods & research*, *44*(2), 186–221.

Caravita, S. C., Sijtsema, J. J., Rambaran, J. A. & Gini, G. (2014). Peer influences on moral disengagement in late childhood and early adolescence. *Journal of youth and adolescence*, *43*(2), 193–207.

Cardwell, S. M., Piquero, A. R., Jennings, W. G., Copes, H., Schubert, C. A. & Mulvey, E. P. (2015). Variability in moral disengagement and its relation to offending in a sample of serious youthful offenders. *Criminal justice and behavior*, *42*(8), 819–839.

Carley, K. (1991). A theory of group stability. *American sociological review*, 331–354.

Cass, R. (2007). Republic. com 2, o.

Castano, E. (2008). On the perils of glorifying the in-group: Intergroup violence, in-group glorification, and moral disengagement. *Social and Personality Psychology Compass*, *2*(1), 154–170.

Centola, D. & Macy, M. (2007). Complex contagions and the weakness of long ties. *American journal of Sociology*, *113*(3), 702–734.

Chen, M., Chen, C. C. & Sheldon, O. J. (2016). Relaxing moral reasoning to win: How organizational identification relates to unethical pro-organizational behavior. *Journal of Applied Psychology*, *101*(8), 1082.

Christian, J. S. & Ellis, A. P. (2014). The crucial role of turnover intentions in transforming moral disengagement into deviant behavior at work. *Journal of business ethics*, *119*(2), 193–208.

Claybourn, M. (2011). Relationships between moral disengagement, work characteristics and workplace harassment. *Journal of Business Ethics*, *100*(2), 283–301.

Clinard, M. B. & Quinney, R. (1973). Corporate criminal behavior. *Criminal behavior systems: a typology.*

Cohen, A. B. (2003). Religion, likelihood of action, and the morality of mentality. *The International Journal for the Psychology of Religion*, *13*(4), 273–285.

Cohen, T. R., Panter, A. T., Turan, N., Morse, L. & Kim, Y. (2014). Moral character in the workplace. *Journal of personality and social psychology*, *107*(5), 943.

Converse, P. E. (1964). The nature of belief systems in mass publics." in david apter, ed., ideology and discontent. new york: Free press.

Cook, J. & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in cognitive science*, *8*(1), 160–179.

Cooper, J. & Fazio, R. H. (1984). A new look at dissonance theory. *Advances in experimental social psychology* (pp. 229–266). Elsevier.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, *31*(3), 187–276.

Critcher, C. R., Huber, M., Ho, A. K. & Koleva, S. P. (2009). Political orientation and ideological inconsistencies:(dis) comfort with value tradeoffs. *Social Justice Research*, *22*(2-3), 181–205.

Curry, O. S., Chesters, M. J. & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation'with a new questionnaire. *Journal of Research in Personality*, *78*, 106–124.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M. & van der Maas, H. L. (2016). Toward a formalized account of attitudes: The causal attitude network (can) model. *Psychological review*, *123*(1), 2.

Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain.

Darley, J. M. & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20.

de Burgh, W. (1930). Right and good: The contradiction of morality. *Journal of Philosophical Studies*, *5*(20), 582–593.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E. & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

DeLisi, M., Peters, D. J., Dansby, T., Vaughn, M. G., Shook, J. J. & Hochstetler, A. (2014). Dynamics of psychopathy and moral disengagement in the etiology of crime. *Youth Violence and Juvenile Justice*, *12*(4), 295–314.

Deppe, K. D., Gonzalez, F. J., Neiman, J., Pahlke, J., Smith, K., Hibbing, J. R. et al. (2015). Reflective liberals and intuitive conservatives: A look at the cognitive reflection test and ideology.

d'Errico, F. & Banks, W. E. (2013). Identifying mechanisms behind middle paleolithic and middle stone age cultural trajectories. *Current Anthropology*, *54*(S8), S371–S387.

d'Errico, F., Banks, W. E. & Clobert, J. (2012). Human expansion: Research tools, evidence, mechanisms. *Dispersal Ecology and Evolution*, *433*.

DeScioli, P., Gilbert, S. S. & Kurzban, R. (2012). Indelible victims and persistent punishers in moral cognition. *Psychological Inquiry*, *23*(2), 143–149.

Detert, J. R., Treviño, L. K. & Sweitzer, V. L. (2008). Moral disengagement in ethical decision making: A study of antecedents and outcomes. *Journal of Applied Psychology*, *93*(2), 374.

Ditto, P. H. & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of personality and social psychology*, *63*(4), 568.

Druckman, J. N. & Bolsen, T. (2011). Framing, motivated reasoning, and opinions about emergent technologies. *Journal of Communication*, *61*(4), 659–688.

Druckman, J. N. & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, *9*(2), 111–119.

Dubois, D. & Prade, H. (1991). Epistemic entrenchment and possibilistic logic. *Artificial Intelligence*, *50*(2), 223–239.

Duffy, M. K., Scott, K. L., Shaw, J. D., Tepper, B. J. & Aquino, K. (2012). A social context model of envy and social undermining. *Academy of management Journal*, *55*(3), 643–666.

Eady, G., Nagler, J., Guess, A., Zilinsky, J. & Tucker, J. A. (2019). How many people live in political bubbles on social media? evidence from linked survey and twitter data. *Sage Open*, *9*(1), 2158244019832705.

Edelmann, A., Wolff, T., Montagne, D. & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, *46*.

Efron, B. et al. (2007). Size, power and false discovery rates. *The Annals of Statistics*, *35*(4), 1351–1377.

Eidelman, S., Crandall, C. S., Goodman, J. A. & Blanchar, J. C. (2012). Low-effort thought promotes political conservatism. *Personality and Social Psychology Bulletin*, *38*(6), 808–820.

Emler, N., Renwick, S. & Malone, B. (1983). The relationship between moral reasoning and political orientation. *Journal of personality and social psychology*, *45*(5), 1073.

Etz, A. & Vandekerckhove, J. (2018). Introduction to bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34.

Evans, J. A. & Aceves, P. (2016). Machine translation: Mining text for social theory. *Annual Review of Sociology*, *42*, 21–50.

Farasat, A., Nikolaev, A., Srihari, S. N. & Blair, R. H. (2015). Probabilistic graphical models in modern social network analysis. *Social Network Analysis and Mining*, *5*(1), 62.

Farrell, J. (2015). Politics: Echo chambers and false certainty. *Nature Climate Change*, *5*(8), 719–720.

Fassin, D. (2012). *A companion to moral anthropology*. John Wiley & Sons.

Feldman, S. & Johnston, C. (2014). Understanding the determinants of political ideology: Implications of structural complexity. *Political Psychology*, *35*(3), 337–358.

Ferreira, J. F., Castelo-Branco, M. & Dias, J. (2012). A hierarchical bayesian framework for multimodal active perception. *Adaptive Behavior*, *20*(3), 172–190.

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.

Festinger, L. (1962). Cognitive dissonance. *Scientific American*, *207*(4), 93–106.

Festinger, L. & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The journal of abnormal and social psychology*, *58*(2), 203.

Fida, R., Paciello, M., Tramontano, C., Fontaine, R. G., Barbaranelli, C. & Farnese, M. L. (2015). An integrative approach to understanding counterproductive work behavior: The roles of stressors, negative emotions, and moral disengagement. *Journal of business ethics*, *130*(1), 131–144.

Fiorina, M. P. (2002). Parties and partisanship: A 40-year retrospective. *Political Behavior*, *24*(2), 93–115.

Fiske, S. T. (2018). *Social beings: Core motives in social psychology.* John Wiley & Sons.

Fränken, J.-P. & Pilditch, T. (2020). Cascades across networks are sufficient for the formation of echo chambers: An agent-based model.

Fränken, J.-P., Theodoropoulos, N. C., Moore, A. B. & Bramley, N. R. (2020). Belief revision in a micro-social network: Modeling sensitivity to statistical dependencies in social learning.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states.* The MIT press.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, *14*(2), 265–285.

Gasse, M. (2017). *Probabilistic graphical model structure learning: Application to multi-label classification* (Doctoral dissertation).

Gerring, J. (1997). Ideology: A definitional analysis. *Political Research Quarterly*, *50*(4), 957–994.

Gini, G., Pozzoli, T. & Hauser, M. (2011). Bullies have enhanced moral competence to judge relative to victims, but lack moral compassion. *Personality and Individual Differences*, *50*(5), 603–608.

Gini, G., Pozzoli, T. & Hymel, S. (2014). Moral disengagement among children and youth: A meta-analytic review of links to aggressive behavior. *Aggressive behavior*, *40*(1), 56–68.

Google. (2019). How google search works — overview. Retrieved February 1, 2021, from https://www.google.com/intl/en_uk/search/howsearchworks/

Gopnik, A., Griffiths, T. L. & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*(2), 87–92.

Graham, J., Haidt, J. & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S. & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Gray, K. & Keeney, J. E. (2015a). Disconfirming moral foundations theory on its own terms: Reply to graham (2015). *Social Psychological and Personality Science*, *6*(8), 874–877.

Gray, K. & Keeney, J. E. (2015b). Impure or just weird? scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science*, *6*(8), 859–868.

Gray, K., Schein, C. & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*(4), 1600.

Gray, K., Waytz, A. & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, *23*(2), 206–215.

Gray, K., Young, L. & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological inquiry*, *23*(2), 101–124.

Greene, J. D. (2008). The secret joke of kant's soul. *Moral psychology*, *3*, 35–79.

Grimm, V. & Railsback, S. F. (2005). *Individual-based modeling and ecology* (Vol. 8). Princeton university press.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, *363*(6425), 374–378.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion.* Vintage.

Haidt, J. (2018, April 20).

Haidt, J. & Bjorklund, F. (2008). Social intuitionists answer six questions about morality.

Haidt, J. & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116.

Haidt, J. & Graham, J. (2009). Planet of the durkheimians, where community, authority, and sacredness are foundations of morality. *Social and psychological bases of ideology and system justification*, 371–401.

Haidt, J., Graham, J. & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2-3), 110–119.

Haidt, J. & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66.

Haidt, J. & Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The innate mind*, *3*, 367–391.

Haidt, J. & Joseph, C. (2011). How moral foundations theory succeeded in building on sand: A response to suhler and churchland. *Journal of Cognitive Neuroscience*, *23*(9), 2117–2122.

Harmon-Jones, E., Harmon-Jones, C. & Levy, N. (2015). An action-based model of cognitive-dissonance processes. *Current Directions in Psychological Science*, *24*(3), 184–189.

Harmon-Jones, E. & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory.

Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong.* Ecco/HarperCollins Publishers.

Helbing, D., Farkas, I. & Vicsek, T. (2000). Simulating dynamical features of escape panic. *Nature*, *407*(6803), 487–490.

Henry, A. D., Prałat, P. & Zhang, C.-Q. (2011). Emergence of segregation in evolving social networks. *Proceedings of the National Academy of Sciences*, *108*(21), 8605–8610.

Hinrichs, K. T., Wang, L., Hinrichs, A. T. & Romero, E. J. (2012). Moral disengagement through displacement of responsibility: The role of leadership beliefs. *Journal of Applied Social Psychology*, *42*(1), 62–80.

Hodge, K., Hargreaves, E. A., Gerrard, D. & Lonsdale, C. (2013). Psychological mechanisms underlying doping attitudes in sport: Motivation and moral disengagement. *Journal of Sport and Exercise Psychology*, *35*(4), 419–432.

Hodge, K. & Lonsdale, C. (2011). Prosocial and antisocial behavior in sport: The role of coaching style, autonomous vs. controlled motivation, and moral disengagement. *Journal of sport and exercise psychology*, *33*(4), 527–547.

Horne, Z., Powell, D., Hummel, J. E. & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112*(33), 10321–10324.

Horne, Z., Powell, D. & Spino, J. (2013). Belief updating in moral dilemmas. *Review of Philosophy and Psychology*, *4*(4), 705–714.

Huang, G.-h., Wellman, N., Ashford, S. J., Lee, C. & Wang, L. (2017). Deviance and exit: The organizational costs of job insecurity and moral disengagement. *Journal of Applied Psychology*, *102*(1), 26.

ICantBreathe. (2020). I can't breathe. Retrieved June 22, 2020, from https://www.facebook.com/hashtag/icantbreathe

IFLScience. (2020). Iflscience. Retrieved June 22, 2020, from https://www.facebook.com/IFLScience/posts/3705594169461567

Inbar, Y., Pizarro, D. A. & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and emotion*, *23*(4), 714–725.

Inbar, Y., Pizarro, D. A. & Bloom, P. (2012). Disgusting smells cause decreased liking of gay men. *Emotion*, *12*(1), 23.

Iyer, R., Koleva, S., Graham, J., Ditto, P. & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS one*, *7*(8), e42366.

Jackson, L. E. & Gaertner, L. (2010). Mechanisms of moral disengagement and their differential use by right-wing authoritarianism and social dominance orientation in support of war. *Aggressive behavior*, *36*(4), 238–250.

Janoff-Bulman, R. & Carnes, N. C. (2013). Surveying the moral landscape: Moral motives and group-based moralities. *Personality and Social Psychology Review*, *17*(3), 219–236.

Jern, A., Chang, K.-m. & Kemp, C. (2009). Bayesian belief polarization. *Advances in neural information processing systems*, 853–861.

Jern, A., Chang, K.-M. K. & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological review*, *121*(2), 206.

Johnson, J. F. & Buckley, M. R. (2015). Multi-level organizational moral disengagement: Directions for future investigation. *Journal of Business Ethics*, *130*(2), 291–300.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P. & Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 2946–2954.

Jones, E. E. (1985). Major developments in social psychology during the past five decades. *Handbook of social psychology*, *1*, 47–107.

Jost, J. T. (2006). The end of the end of ideology. *American psychologist*, *61*(7), 651.

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political psychology*, *38*(2), 167–208.

Jost, J. T., Federico, C. M. & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual review of psychology*, *60*, 307–337.

Juslin, P., Nilsson, H., Winman, A. & Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition*, *120*(2), 248–267.

Kahan, D. M. (2007). The cognitively illiberal state. *Stan. L. Rev.*, *60*, 115.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D. & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature climate change*, *2*(10), 732–735.

Kelman, H. C. & Hamilton, V. L. (1989). *Crimes of obedience: Toward a social psychology of authority and responsibility.* Yale University Press.

Keltner, D. & Robinson, R. J. (1996). Extremism, power, and the imagined basis of social conflict. *Current Directions in Psychological Science*, *5*(4), 101–105.

Kennedy, J. A., Kray, L. J. & Ku, G. (2017). A social-cognitive approach to understanding gender differences in negotiator ethics: The role of moral identity. *Organizational Behavior and Human Decision Processes*, *138*, 28–44.

Kertzer, J. D., Powers, K. E., Rathbun, B. C. & Iyer, R. (2014). Moral support: How moral values shape foreign policy attitudes. *The Journal of Politics*, *76*(3), 825–840.

Killen, M., Smetana, J. et al. (2005). *Handbook of moral development.* Psychology Press.

Kinder, D. R. & Kalmoe, N. P. (2017). *Neither liberal nor conservative: Ideological innocence in the american public.* University of Chicago Press.

Kinoshita, S. & Norris, D. (2012). Task-dependent masked priming effects in visual word recognition. *Frontiers in Psychology*, *3*, 178.

Klayman, J. & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, *94*(2), 211.

Knobloch-Westerwick, S., Mothes, C. & Polavin, N. (2020). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research*, *47*(1), 104–124.

Knoll, M., Lord, R. G., Petersen, L.-E. & Weigelt, O. (2016). Examining the moral grey zone: The role of moral disengagement, authenticity, and situational strength in predicting unethical managerial behavior. *Journal of Applied Social Psychology*, *46*(1), 65–78.

Knutsen, O. (1995). Value orientations, political conflicts and left-right identification: A comparative study. *European journal of political research*, *28*(1), 63–93.

Koelle, D., Pfautz, J., Farry, M., Cox, Z., Catto, G. & Campolongo, J. (2006). Applications of bayesian belief networks in social network analysis. *Proceedings of the 4th Bayesian modeling applications workshop during the 22nd annual conference on uncertainty in artificial intelligence.*

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. *Handbook of socialization theory and research*, *347*, 480.

Koleva, S. & Haidt, J. (2012). Let's use einstein's safety razor, not occam's swiss army knife or occam's chainsaw. *Psychological Inquiry*, *23*(2), 175–178.

Koller, D. & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT press.

Krosnick, J. A. & Alwin, D. F. (1989). Aging and susceptibility to attitude change. *Journal of personality and social psychology*, *57*(3), 416.

Kugler, M., Jost, J. T. & Noorbaloochi, S. (2014). Another look at moral foundations theory: Do authoritarianism and social dominance orientation explain liberal-conservative differences in "moral" intuitions? *Social Justice Research*, *27*(4), 413–431.

LADbible. (2019). Dogs and cats can get along. Retrieved June 22, 2020, from https://www.facebook.com/199098633470668/videos/2393040087614505/

Lakoff, G. (2002). Moral politics: How liberals and conservatives think . chicago, il, us.

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Lazarsfeld, P. F., Berelson, B. & Gaudet, H. (1944). The people's choice.

Lee, A., Schwarz, G., Newman, A. & Legood, A. (2019). Investigating when and why psychological entitlement predicts unethical pro-organizational behavior. *Journal of Business Ethics*, *154*(1), 109–126.

Lee, K., Kim, E., Bhave, D. P. & Duffy, M. K. (2016). Why victims of undermining at work become perpetrators of undermining: An integrative model. *Journal of Applied Psychology*, *101*(6), 915.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7.

Leidner, B., Castano, E., Zaiser, E. & Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin*, *36*(8), 1115–1129.

Leifeld, P. (2014). Polarization of coalitions in an agent-based model of political discourse. *Computational Social Networks*, *1*(1), 7.

Levin, J. (2006). Choice under uncertainty. *Lecture Notes*, *2*.

Lewandowsky, S. & Oberauer, K. (2016). Motivated rejection of science. *Current Directions in Psychological Science*, *25*(4), 217–222.

Lewicka, M. (1988). On objective and subjective anchoring of cognitive acts: How behavioural valence modifies reasoning schemata. *Recent trends in theoretical psychology* (pp. 285–301). Springer.

Lewicka, M. (1998). Confirmation bias. *Personal control in action* (pp. 233–258). Springer.

Lewis, G. J. & Bates, T. C. (2011). From left to right: How the personality system allows basic traits to influence politics via characteristic moral adaptations. *British Journal of Psychology*, *102*(3), 546–558.

Lichtblau, E. (2020). Doj memo: Armed agents can be deployed at vote counting. Retrieved November 19, 2020, from https://time.com/5907600/doj-armed-agents-polls/

Lim, X. J., Radzol, A. M., Cheah, J. & Wong, M. (2017). The impact of social media influencers on purchase intention and the mediation effect of customer attitude. *Asian Journal of Business Research*, *7*(2), 19–36.

Lodge, M. & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

Lord, C. G., Ross, L. & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, *37*(11), 2098.

Macy, M. W. (2015). An emerging trend: Is big data the end of theory? *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, 1–14.

Macy, M. W. & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual review of sociology*, *28*(1), 143–166.

Madsen, J. K., Bailey, R., Carrella, E. & Koralus, P. (2019). Analytic versus computational cognitive models: Agent-based modeling as a tool in cognitive sciences. *Current Directions in Psychological Science*, *28*(3), 299–305.

Madsen, J. K., Bailey, R. & Pilditch, T. D. (2017). Growing a bayesian conspiracy theorist: An agent-based model.

Madsen, J. K., Bailey, R. M. & Pilditch, T. D. (2018). Large networks of rational agents form persistent echo chambers. *Scientific reports*, *8*(1), 12391.

Malka, A. & Lelkes, Y. (2010). More than ideology: Conservative–liberal identity and receptivity to political cues. *Social Justice Research*, *23*(2-3), 156–188.

Mandel, D. R. (2014). The psychology of bayesian reasoning. *Frontiers in Psychology*, *5*, 1144.

Martin, S. R., Kish-Gephart, J. J. & Detert, J. R. (2014). Blind forces: Ethical infrastructures and moral disengagement in organizations. *Organizational Psychology Review*, *4*(4), 295–325.

McAlister, A. L., Bandura, A. & Owen, S. V. (2006). Mechanisms of moral disengagement in support of military force: The impact of sept. 11. *Journal of Social and Clinical Psychology*, *25*(2), 141–165.

McFarland, D. A., Lewis, K. & Goldberg, A. (2016). Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist*, *47*(1), 12–35.

McFerran, B., Aquino, K. & Duffy, M. (2010). How personality and moral identity relate to individuals' ethical ideology. *Business Ethics Quarterly*, 35–56.

Mercier, H. & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory.

Mikhail, J. M. (2000). *Rawls' linguistic analogy: A study of the" generative grammar" model of moral theory described by john rawls in" a theory of justice".* (Doctoral dissertation). ProQuest Information & Learning.

Miles, A. & Vaisey, S. (2015). Morality and politics: Comparing alternate theories. *Social Science Research*, *53*, 252–269.

Miller, D. T. (2021). Characterizing qanon: Analysis of youtube comments presents new conclusions about a popular conservative conspiracy. *First Monday*.

Molina, M. & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*.

Moore, A., Hong, S. & Cram, L. (2021). Trust in information, political identity and the brain: An interdisciplinary fmri study. *Philosophical Transactions of the Royal Society B*, *376*(1822), 20200140.

Moore, C. (2008). Moral disengagement in processes of organizational corruption. *Journal of Business ethics*, *80*(1), 129–139.

Moore, C., Detert, J. R., Klebe Treviño, L., Baker, V. L. & Mayer, D. M. (2012). Why employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychology*, *65*(1), 1–48.

Mutz, D. C. & Martin, P. S. (2001). Facilitating communication across lines of political difference: The role of mass media. *American political science review*, 97–114.

Neal, Z. P. (2020). A sign of the times? weak and strong polarization in the us congress, 1973–2016. *Social Networks*, *60*, 103–112.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42.

Ngampruetikorn, V. & Stephens, G. J. (2016). Bias, belief, and consensus: Collective opinion formation on fluctuating networks. *Physical Review E*, *94*(5), 052312.

Oberauer, K. (2018). From words to models: Building a toolkit. In S. Farrel & S. Lewandowsky (Eds.), *Computational modeling of cognition and behavior* (pp. 40–43). Cambridge University Press.

Obermann, M.-L. (2013). Temporal aspects of moral disengagement in school bullying: Crystallization or escalation? *Journal of school violence*, *12*(2), 193–210.

Ogunfowora, B. & Bourdage, J. S. (2014). Does honesty–humility influence evaluations of leadership emergence? the mediating role of moral disengagement. *Personality and Individual Differences*, *56*, 95–99.

Ogunfowora, B., Bourdage, J. S. & Nguyen, B. (2013). An exploration of the dishonest side of self-monitoring: Links to moral disengagement and unethical business decision making. *European Journal of Personality*, *27*(6), 532–544.

Osofsky, M. J., Bandura, A. & Zimbardo, P. G. (2005). The role of moral disengagement in the execution process. *Law and Human Behavior*, *29*(4), 371–393.

Pancs, R. & Vriend, N. J. (2007). Schelling's spatial proximity model of segregation revisited. *Journal of Public Economics*, *91*(1-2), 1–24.

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you.* Penguin UK.

Parry, H. R. & Bithell, M. (2012). Large scale agent-based modelling: A review and guidelines for model scaling. *Agent-based models of geographical systems* (pp. 271–308). Springer.

Pennycook, G. & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Pentland, A. (2015). *Social physics: How social networks can make us smarter.* Penguin.

Petersen, M. B., Skov, M., Serritzlew, S. & Ramsøy, T. (2013). Motivated reasoning and political parties: Evidence for increased processing in the face of party cues. *Political Behavior*, *35*(4), 831–854.

Piaget, J. (1965). The stages of the intellectual development of the child. *Educational psychology in context: Readings for future teachers*, *63*(4), 98–106.

Pilditch, T. D. (2017). Opinion cascades and echo-chambers in online networks: A proof of concept agent-based model.

Pornari, C. D. & Wood, J. (2010). Peer and cyber aggression in secondary school students: The role of moral disengagement, hostile attribution bias, and outcome expectancies. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *36*(2), 81–94.

Pratto, F., Sidanius, J. & Levin, S. (2006). Social dominance theory and the dynamics of intergroup relations: Taking stock and looking forward. *European review of social psychology*, *17*(1), 271–320.

Pratto, F., Sidanius, J., Stallworth, L. M. & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of personality and social psychology*, *67*(4), 741.

Prinz, J. (2007). *The emotional construction of morals*. Oxford University Press.

Quine, W. V. (1951). Main trends in recent philosophy: Two dogmas of empiricism. *The philosophical review*, 20–43.

Rai, T. S. & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, *118*(1), 57.

Read, S. J. & Monroe, B. M. (2007). Modeling cognitive dissonance using a recurrent neural network model with learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*(29).

Reuband, K.-H. (1991). Moral beliefs: Patterns of crystallization and individual stability. findings from a panel study. s. 573-580 in: Albrecht, g./otto, h.-u.(hrsg.), social prevention. theoretical controversies, research problems, and evaluation strategies.

Roccato, M. & Ricolfi, L. (2005). On the correlation between right-wing authoritarianism and social dominance orientation. *Basic and applied social psychology*, *27*(3), 187–200.

Rosenberg, M. J., Hovland, C. I., McGuire, W. J., Abelson, R. P. & Brehm, J. W. (1960). Attitude organization and change: An analysis of consistency among attitude components.(yales studies in attitude and communication.), vol. iii.

Sakai, H. (1999). A multiplicative power-function model of cognitive dissonance: Toward an integrated theory of cognition, emotion, and behavior after leon festinger. *Convention of the Japanese Psychological Association, 48th, Oct, 1984, Osaka, Japan; Portions of this chapter were presented at the 48th Convention of the Japanese Psychological Association, Osaka, Japan, Oct 1984, and at the 7th International Kurt Lewin Conference, Los Angeles, California, Sep 1996.*

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age.* Princeton University Press.

Samnani, A.-K., Salamon, S. D. & Singh, P. (2014). Negative affect and counterproductive workplace behavior: The moderating role of moral disengagement and gender. *Journal of business ethics*, *119*(2), 235–244.

Scheibehenne, B. & Studer, B. (2014). A hierarchical bayesian model of the influence of run length on sequential predictions. *Psychonomic bulletin & review*, *21*(1), 211–217.

Schein, C. & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, *41*(8), 1147–1163.

Schein, C. & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70.

Schelling, T. C. (1969). Models of segregation. *The American Economic Review*, *59*(2), 488–493.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, *1*(2), 143–186.

Schrager, L. S. & Short Jr, J. F. (1978). Toward a sociology of organizational crime. *Social problems*, *25*(4), 407–419.

Schuldt, J. P., Pearson, A. R., Romero-Canyas, R. & Larson-Konar, D. (2017). Brief exposure to pope francis heightens moral beliefs about climate change. *Climatic Change*, *141*(2), 167–177.

Sears, D. O. & Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, *31*(2), 194–213.

Seliger, M. (1976). Politics and ideology. *Ruskin House, London.*

Shultz, T. R. & Lepper, M. R. (1992). A constraint satisfaction model of cognitive dissonance phenomena. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 462–467.

Shultz, T. R. & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological review*, *103*(2), 219.

Shultz, T. R. & Lepper, M. R. (1998). The consonance model of dissonance reduction. *Connectionist models of social reasoning and social behavior*, 211–244.

Shultz, T. R., Léveillé, E. & Lepper, M. R. (1999). Free choice and cognitive dissonance revisited: Choosing "lesser evils" versus "greater goods". *Personality and Social Psychology Bulletin*, *25*(1), 40–48.

Shweder, R. A. & Haidt, J. (1993). The future of moral psychology: Truth, intuition, and the pluralist way. *Psychological science*, *4*(6), 360–365.

Shweder, R. A., Much, N. C., Mahapatra, M. & Park, L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. *Morality and health*, *119*, 119–169.

Sidanius, J. & Pratto, F. (2001). *Social dominance: An intergroup theory of social hierarchy and oppression.* Cambridge University Press.

Sidanius, J., Pratto, F. & Bobo, L. (1994). Social dominance orientation and the political psychology of gender: A case of invariance? *Journal of Personality and Social Psychology*, *67*(6), 998.

Siddharth, N., Paige, B., Van de Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F. & Torr, P. (2017). Learning disentangled representations with semisupervised deep generative models. *Advances in Neural Information Processing Systems*, 5925–5935.

Sili, A., Fida, R., Zaghini, F., Tramontano, C. & Paciello, M. (2014). Counterproductive behaviors and moral disengagement of nurses as potential consequences of stress-related work: Validity and reliability of measurement scales. *La Medicina del lavoro*, *105*(5), 382–394.

Simpson, A. (n.d.). Moral foundations theory: Background, review, and scaffolding for future research. *Encyclopedia of Personality and Individual Differences*, 1–19.

Sindermann, C., Elhai, J. D., Moshagen, M. & Montag, C. (2020). Age, gender, personality, ideological attitudes and individual differences in a person's news spectrum: How many and who might be prone to "filter bubbles" and "echo chambers" online? *Heliyon*, *6*(1), e03214.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, *4*(4), 267–281.

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, *34*(3), 150–160.

Starnini, M., Frasca, M. & Baronchelli, A. (2016). Emergence of metapopulations and echo chambers in mobile agents. *Scientific reports*, *6*, 31834.

Stevens, G. W., Deuling, J. K. & Armenakis, A. A. (2012). Successful psychopaths: Are they unethical decision-makers and why? *Journal of Business Ethics*, *105*(2), 139–149.

Steyvers, M., Lee, M. D. & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Suhler, C. L. & Churchland, P. (2011). Can innate, modular "foundations" explain morality? challenges for haidt's moral foundations theory. *Journal of cognitive neuroscience*, *23*(9), 2103–2116.

Sunstein, C. R. (2001a). *Echo chambers: Bush v. gore, impeachment, and beyond.* Princeton University Press Princeton, NJ.

Sunstein, C. R. (2001b). *Republic. com.* Princeton university press.

Szwajkowski, E. (1985). Organizational illegality: Theoretical integration and illustrative application. *Academy of Management Review*, *10*(3), 558–567.

Taber, C. S. & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, *50*(3), 755–769.

Tasa, K. & Bell, C. M. (2017). Effects of implicit negotiation beliefs and moral disengagement on negotiator attitudes and deceptive behavior. *Journal of business ethics*, *142*(1), 169–183.

Tewksbury, D. (2005). The seeds of audience fragmentation: Specialization in the use of online news sites. *Journal of broadcasting & electronic media*, *49*(3), 332–348.

Tjosvold, D. & Sun, H. F. (2002). Understanding conflict avoidance: Relationship, motivations, actions, and consequences. *International Journal of Conflict Management*, *13*(2).

Traveller. (2020). Traveller. Retrieved June 22, 2020, from https://www.facebook.com/TravellerAU/posts/3010143405770883

Turiel, E. (1983). *The development of social knowledge: Morality and convention.* Cambridge University Press.

Turner-Zwinkels, F. M., Johnson, B. B., Sibley, C. G. & Brandt, M. J. (2020). Conservatives' moral foundations are more densely connected than liberals' moral foundations. *Personality and Social Psychology Bulletin*, 0146167220916070.

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M. & Depaoli, S. (2017). A systematic review of bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217.

Van Overwalle, F. & Jordens, K. (2002). An adaptive connectionist model of cognitive dissonance. *Personality and Social Psychology Review*, *6*(3), 204–231.

Verhulst, B., Eaves, L. J. & Hatemi, P. K. (2012). Correlation not causation: The relationship between personality traits and political ideologies. *American journal of political science*, *56*(1), 34–51.

Villegas de Posana, C., Florez, J. & Espinel, N. (2018). Moral disengagement mechanisms and armed violence. a comparative study of paramilitaries and guerrillas in colombia. *Revista Colombiana de Psicologia*, *27*(1), 55–69.

Vitell, S. J., Keith, M. & Mathur, M. (2011). Antecedents to the justification of norm violating behavior among business practitioners. *Journal of Business ethics*, *101*(1), 163–173.

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K. & Storms, G. (2015). How do people learn from negative evidence? non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1–25.

Voorspoels, W., Storms, G. & Vanpaemel, W. (2012). Contrast effects in typicality judgements: A hierarchical bayesian approach. *Quarterly Journal of Experimental Psychology*, *65*(9), 1721–1739.

Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of sociology*, *105*(2), 493–527.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, *99*(9), 5766–5771.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*(1), 1–5.

Waytz, A. & Epley, N. (2012). Social connection enables dehumanization. *Journal of experimental social psychology*, *48*(1), 70–76.

Weber, C. R. & Federico, C. M. (2013). Moral foundations and heterogeneity in ideological preferences. *Political Psychology*, *34*(1), 107–126.

Whalen, A., Griffiths, T. L. & Buchsbaum, D. (2018). Sensitivity to shared information in social learning. *Cognitive science*, *42*(1), 168–187.

Wikipedia. (2020). Revolutionary armed forces of colombia — Wikipedia, the free encyclopedia [[Online; accessed 22-December-2020]]. https://en.wikipedia.org/wiki/Revolutionary_Armed_Forces_of_Colombia

Williams Jr, R. M. (1979). Change and stability in values and value systems: A sociological perspective. *Understanding human values*, *15*, 46.

Wojcik, S. P., Hovasapian, A., Graham, J., Motyl, M. & Ditto, P. H. (2015). Conservatives report, but liberals display, greater happiness. *Science*, *347*(6227), 1243–1246.

Yang, Y., Roux, A. V. D., Auchincloss, A. H., Rodriguez, D. A. & Brown, D. G. (2011). A spatial agent-based model for the simulation of adults' daily walking within a city. *American journal of preventive medicine*, *40*(3), 353–361.

Yılmaz, O. & Sarıbay, S. A. (2016). An attempt to clarify the link between cognitive style and political ideology: A non-western replication and extension.

Zhang, J. (2004). Residential segregation in an all-integrationist world. *Journal of Economic Behavior & Organization*, *54*(4), 533–550.

Zhao, H., Zhang, H. & Xu, Y. (2019). Effects of perceived descriptive norms on corrupt intention: The mediating role of moral disengagement. *International Journal of Psychology*, *54*(1), 93–101.

Zheng, X., Qin, X., Liu, X. & Liao, H. (2019). Will creative employees always make trouble? investigating the roles of moral identity and moral disengagement. *Journal of Business Ethics*, 1–20.

Zhou, Y. (2011). Structure learning of probabilistic graphical models: A comprehensive survey. *arXiv preprint arXiv:1111.6925.*

Zimmerman, J. L. & Reyna, C. (2013). The meaning and role of ideology in system justification and resistance for high-and low-status people. *Journal of Personality and Social Psychology*, *105*(1), 1.