Structure is Management of Uncertainty in Groups

Jesse Hoey
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario,
Canada, N2L3G1
jhoey@cs.uwaterloo.ca

January 4, 2021

Abstract

I argue that the management of uncertainty by agents in a social world is foundational to the formation of social structures and to the definition of culture. I present a deep Bayesian model for this management of uncertainty in intelligent systems, and I argue for its applicability to cultural sociology. As social systems grow more heterogeneous, management of uncertainty in any participating agent becomes computationally difficult, and I propose that combinations of a small number of layers of reasoning in a deep Bayesian model are sufficient to account for some of the salient ways by which humans manage this uncertainty. Three forces come into play when considering such a model, and each is connected to a particular form of uncertainty. A *denotative* layer in the model represents uncertainty in the world or environment (ambiguity and risk about outcomes), a *connotative layer* manages the uncertainty about relationships with other social agents, and the *connection* between denotative and connotative handles uncertainty about identities of the self and others. Behaviours taken by agent and by others are handled in both layers simultaneously. I show how the tradeoff between these three factors maps to different social structures, and I use the model to make predictions across a range of domains, and show its relationship to cultural sociological, social psychological, economic and sociological theorizing. I further link this model to Bayesian views of the mind, primarily the active inference model of human intelligence, and compare and contrast to more traditional artificial intelligence.

Preface

It must be noted that I am a computer scientist trained as a physicist, who is interested in better understanding how humans think by building machines that try to think. My current focus is on building machines that can think about human-machine social groups. I was led to this through mathematics. Many of the speculative statements in this paper arise because of the geometry of the parameter space (e.g. “there must be another axis here”) or by the information theoretic properties of the Bayesian model (e.g. another method, theoretically defined by empirically untried, of getting the same outcome.). I therefore must ask you, dear Reader, to forgive these little scientific infractions as exactly speculative, perhaps made more meaningful because of their theoretical background, but largely untested. Students should take these as opportunities for further investigation, not as claims of truth or scientific fact ¹

¹As Professor Mori said in a talk at ICRA 2015 in Toyko: “I just tell students where to dig!”. Mori is renowned for his “uncanny valley” concept (Mori, 1970).

Contents

1	Introduction	4
2	The Management of Uncertainty	9
2.1	Alignment	9
2.2	Deep Bayesian Modeling	14
2.3	Parameter space and the Simplex	15
3	Social-Psychological Perspective	19
3.1	Uncertainty	20
3.2	Mechanical and organic solidarity	22
3.3	Three Freedoms	23
3.4	Self	26
3.5	Freedom and Equality	28
3.6	Security and Rights	29
3.7	An example	29
4	Affect Control Theory and <i>BayesAct</i>	30
4.1	Deep Bayesian Models	33
4.2	Flashlight algorithms	37
4.3	Free Energy and Active Inference	40
5	Social Structures	48
5.1	From structures to settings...	49
5.2	...and back again (from settings to structures)	56
5.3	Smarter Agents	58
6	Psychological Economics and Political Sociology	59
6.1	Political Science	62
6.2	Government, Sovereign and Law	67
6.3	Neutral Point Decisions	68
7	Cultural Sociology	70
7.1	Durkheim	73
7.2	Lévi-Strauss	78
7.3	Bales	79
7.4	White	83

7.5	Martin	86
7.6	Values and Interests	86
7.7	Moreno	89
7.8	Osgood/Heise/MacKinnon	89
7.9	Traditional, Toolkit, and Practice	90
7.10	Lawler	95
7.11	Social Cohesion	96
7.12	Schwartz	97
7.13	Self-Affirmation Theory	100
7.14	Personality Theories	101
7.15	Morality	103
8	Artificial Intelligence	103
8.1	Reinforcement Learning	104
8.2	Relating Uncertainty and Emotion	106
8.3	Ethical Considerations	107
9	Dyanmics of Social Networks	108
9.1	Changing structures	108
9.1.1	Denotative Solution	110
9.1.2	Connotative Solution	111
9.1.3	Final network, Tradeoffs and Mutiny	111
9.2	Conformity	112
9.3	Exogenous events	113
9.4	Ambiguity	116
9.5	Political Polarization	118
9.6	David Heise	121
9.7	Automatic Government	122
10	Conclusion	125

1 Introduction

The philosophical problem of uniting the objective and the subjective in a social world has been the object of millenia of investigation. Idea, or abstract thought, has always been seen as something un-real, and much of western philosophy has been engaged in attempting to prove this arises because of a supreme being (Russell, 1945). Harkening back to Plato, the abstract “idea” of a generic object (say a bed, or a mathematical operation) has an existence of its own, and is “superior” in some way to all such objects (e.g. all individual, real, beds). The existence of this abstract “object”, since it is perfect, must arise from some supreme being, some god. The disengagement from any supreme being in the seventeenth century allowed for a more sophisticated analysis of the nature of knowledge, resulting in its attribution to three factors proposed by Locke: intuition, agreement of ideas, and sensation (Locke, 1690). This is further reflected in Locke’s political philosophy in which he discusses three freedoms defined by three constraints: permission from others, the will of others, and the laws of nature. While the permission from others is held intuitively (I have freedom to act how I see fit without explicitly asking permission, but social contracts restrict the space of actions I know I can take), the will of others must cohere with the actions of the self, restricting another kind of freedom. Finally, the laws of nature are unavoidable and must be adhered to, restricting yet another type of freedom to carry out whatever strategy we desire (e.g. the laws of nature prevent me from doing anything that involves walking on water). Although not investigated in detail by Locke, these three factors are mutually exclusive, in that knowledge can be held according to one at the expense of another. I can force ideas to agree if I don’t worry about contradicting sensations and intuition, for example. My goal in this essay is to show how these three elements of the nature of knowledge (and freedom) are in fact simply different ways of managing uncertainty in the human mind. Uncertainty may arise from the external world, leading to a lack of “sensation” knowledge, or from the internal (mental) world, leading to a lack of “intuitive” knowledge, or from the agreement between the two, leading to a lack of “agreement of ideas”. These three are mutually exclusive, but the exact tradeoff used between the three is arbitrary: a group may function as a society using any of the infinitude of possible “settings” of these three uncertainty management elements, so long as the whole group is using the same settings. I propose that much political strife could be avoided by simply accepting this fact. Once accepted, there is no longer any “right” and “wrong” setting, simply a need to understand, accept, and adapt to groups using different settings. I will base much of my discussion on the fact that these settings result in different structures in human social networks, another long-standing idea in cultural sociology and political philosophy.

Indeed, many disciplines, from sociology to artificial intelligence, have sought to unify the observed structures in human cultural social networks with the individual properties of intelligent agents. The highly complex system of human interrelationships somehow produces what can be observed as a somewhat stable “culture.” It is in this pursuit that I make a connection between these large-scale social structures seen in human social networks and some fundamental properties of the management of uncertainty and information in an intelligent agent. I show that these properties are fundamental in the sense of being based on physical principles of energy and information in any living agent. The oft-referenced “fight” against the second law of thermodynamics is known to be a flawed concept for existence. Instead, these fundamental properties can be viewed as arising from an agent leveraging the second law for its own benefit by amassing

orderly states (information) at the expense of externalities (Jeffery et al., 2019). Such an agent needs to implement a controller that is somehow aligned with its environment. A classic result in control theory, the “best” such controller is a model of the system it is controlling (Conant and Ashby, 1970). Therefore, I define an intelligent agent as being exactly one that implements these fundamental physical properties of energy and information. This definition requires careful consideration before proceeding. The reader may immediately think that I am setting him² up with a tautological basis that I will use later to pull the wool over his eyes. Nothing can be further from the truth! There are quite significant reasons to believe that intelligence as defined according to these energy and information principles (subsumed by the “*free energy principle*”) will be more successful than intelligence not so defined (Friston, 2010; Ramstead et al., 2019). In fact, there is also reason to believe that intelligence *of a group* defined according to these principles will be more successful than a group defined otherwise (Moutoussis et al., 2014b). In the same spirit of a vast body of work in cultural sociology, I work from the hypothesis that there are “structural potentials” (Martin, 2009, p.x) that are inherent in the ways in which certain agents maintain their relationships, and vice-versa, that there are certain relationships that result in the revelation of social structures.

It is the management of uncertainty which plays a critical role in the success of an agent. Uncertainty arises because of environmental noise,³ and because of internal (to the agent) dispersion due to the natural loss of information in a noisy sequential system. It is the “matching” of these uncertainties that an agent is most concerned with, and there is evidence that deep models are important for this matching process to occur in complex (e.g. social) environments. An agent embedded in a social network has an even more troubling issue, and that is that his environment comprises other intelligent agents with whom he must cooperate in order to survive (or succeed). As the complexity of this system grows very quickly with the number of agents added, the modeling capacity of the agent comes into play. When this capacity is reached, any further additions to the social network would induce a corresponding increase in dispersion of the agent’s model. This increase corresponds in turn to a decreased performance as its predictions are less precise. I therefore show that a second, deeper, layer of representation can somewhat resolve this problem, by constructing an “abstract” version of the parts of the (overstretched) current model I have been describing. This abstraction describes the behaviours of others which is the largest component of the model by far.

The reader will wonder what I have gained by doing this. Other agents are likely to be one of the more difficult to predict elements of an agent’s environment, and one could presumably gain by engaging in this more complex modeling (notwithstanding the resource limitations). In certain restricted situations (e.g. a well known individual in a highly predictable environment) this may be possible, but will quickly become intractable beyond a small number. However, with two assumptions we can make progress. First, let us suppose that all agents in a group *share* an abstract model of their social econiche to some reasonable degree of approximation.⁴ Second, let us endow agents with the ability to *communicate* some signals that refer *directly* to

²I use “him” and “her” indiscriminately and without prejudice in this paper, and either, or “they” may be substituted throughout without altering the thesis presented.

³I use noise here for environment to please the engineers, but will use dispersion throughout to please the scientists and mathematicians. Disperse distributions are noisy, undispersed ones are not.

⁴This type of sharing is often referred to as “common knowledge”, in contrast to “shared knowledge” which is usually defined as between only two agents. Common knowledge is held by all members of a group, and all members know that (Ridgeway, 2019).

this abstract model. With these two assumptions in hand, we can argue that an agent using this abstraction as a guide for behaviour will be more predictable by his peers, and his peers will be more predictable to him. Therefore, they can jointly engage in cooperative solutions to social dilemmas. Agents that are not well aligned can use the signaling mechanism to relay differentials between the way the interaction is going, and the way they think it should be going. These differentials reveal the shifts in the abstract space necessary to regain alignment, usually referred to as “emotional” signals (e.g. facial expressions).

Given that all members of the social grouping of the agent share the same abstract model, one would expect all agents to have their overstretched models relieved of this burden, allowing the interactions with other humans to be handled at a more abstract level. We are somewhat used to referring to this abstract model as “emotional” while the remaining pieces are “cognitive”, but here I will avoid these terms and will use “connotative” (meaning, emotions and sentiments) and “denotative” (being, deliberation and reason). Once such a deep model is constructed, one has the difficult job of setting the parameters of it, for both denotative and connotative states. This job is however, taken care of through the Bayesian statistics, and could be further the subject of an even higher level in the model. Nevertheless, the final setting of these parameters will change the behaviours of the agent, as its predictions of its own actions will change.

An agent would be pressured into doing this separation into a deep model for a good reason. Doing so, and ensuring a shared abstract model between a given group, would allow the group to “guide” all members simultaneously by shifting sentiments slightly. In fact, this happens naturally as an agent joins a group. That is, any intelligent agent, when embedded in a group of other intelligent agents, will adopt its parametric model to better fit that of its environment (both its social group, itself, and its physical environment). While the parameter settings are somewhat arbitrary, exactly *how* this fit is obtained will lead to substantially different behaviours on the parts of agents involved. Further we can see that the fit may be obtained both by moving the new individual towards the group, and by moving the group towards the individual. It is likely that this operates like a gravitational field, in that larger groups are less affected (“pulled”) by the new individual.

However, for a social group and each individual agent to function properly, the group needs to “share” this “fit” with all members. Achieving this requires a matching of the tradeoffs between the two layers in the deep Bayesian model. This matching, as I will show, can be reliably done in a group only for certain, restricted, settings of the parameters. Stepping outside of this “restricted” zone is possible, certainly more difficult, and potentially more dangerous. Within it, however, there may be subtle differences in the ways in which uncertainty is “shared” between the different layers in the deep Bayesian model. These subtle differences, in fact, turn out to be precisely those that lead to certain social structures forming between the individuals. Therefore, I expect to find groups which are much more reliant on the abstract “emotional” side of an interaction, while others who are more reliant on the concrete side of it. The former lead to transitive and diverse groups, the latter to isolated and homogeneous groups.

I will present this matching problem using an allegory of two boys searching for each other in the dark with a flashlights, as shown in Figure 1. The flashlights have an adjustable beam width, from narrow and far to wide and short. The boys get rewarded for how much light the other records (the **density** of light falling on him). One can see that for certain settings of flashlight beams, the boys have no hope. If one sets his beam on small and far, but the other does as well, they will have trouble finding each other. If both beams are wide, they can easily find each

Figure 1: Allegorical example to demonstrate the uncertainty matching principle. Boy A (he can be A-narrow or A-wide) and boy B (B-narrow or B-wide) attempt to find each other’s flashlights and are rewarded by the density of light falling on the other boy’s flashlight.

other, but the density of both together is low. Thus, they can either both use medium sized beams, or one can use a small beam and the other a large one.

In the flashlight allegory above, consider the targets for each boy (the other boy) are like the social world, and the flashlight is the boy’s predictions of how the social world will behave on a level of “meaning”. The size of the target is the diversity of the social world, and represents the variance in expected behaviours. The size of the flashlight is the strength of the abstract social model the agent is building (his prior model). Therefore, the “allowable” settings are those that combine high diversity with strong abstract predictions and those that combine low diversity with weak abstract predictions. Both of these settings may work well in a network of agents with the same settings,⁵ but may conflict when put in groups with different settings, as the new agent’s existing model will necessarily break down (and not match). The process of learning the new “fit” will be one that may be individual dependent. Such “spotlight” metaphors have been deeply explored in the context of psychological (usually visual) attention (Crick, 1984).

I will first show how the social problem allegorized by the flashlight narrative has some remarkable similarities with some fundamental information theoretic properties of agents interacting in a group. I then make a hypothesis that these three dimensions of uncertainty management are, in fact, dimensions of freedom, and simultaneously, dimensions of equality. I will argue for this hypothesis using a review of some of the major literature in political science on the subject. I will discuss how only a limited space of settings in this three dimensional space is actually functional, and this limited space forms a two-dimensional manifold lying on a *simplex*

⁵Note that this does not mean necessarily that the agents are homogeneous. While their parameter settings may be the same, the parameter settings define the *space* of possible models and agent can take on, and therefore the parameter settings are more of a measure of a social group’s expansiveness. Granted, boys with wider flashlight beams will have settings that rely less on abstract meanings.

in the middle of the uncertainty space. I then present the deep Bayesian model, *BayesAct*, as a candidate to model the key elements of this simplex in the space of parameters. In Section 2.1, I show how the same simplex-like manifold of social structures arises naturally from some key information theoretic properties of this model,

I then use this model in a number of different ways. I start by showing that the fundamental properties of the management of uncertainty (which turn out to be parameters of the deep Bayesian model) and specific triadic social structures have a strong dependence. I first show, for every possible triadic structure, what type of uncertainty management would be required by a population of resource-bounded agents. I then show, how only certain structures would sustain a population of agents with a specific parameter setting. I demonstrate how these two correspond and therefore how the space of structures is can be used in an equivalent manner to the parameter space in the deep Bayesian model. This is an interesting result because it allows the simulation of arbitrary social structures including the interactions between different social structures (and therefore cultures). Further, it allows the development of practical tools that can be used to help guide social structures through policy implementation.

Having thus established that the deep Bayesian model is an appropriate one for the modeling of social structures, and that it has connections to certain important aspects of political science, I proceed to make an association between this parameter space and the two- and three- dimensional spaces explored in the philosophical and sociological literatures (Rousseau, 1762; Berlin, 2002; Durkheim, 2014/1893; Moreno, 1934; Feld and Elmore, 1982; White, 1992; Schwartz, 1994; Bales, 1999; Martin, 2009; Lawler et al., 2009; Anderson, 2017) and (Ridgeway, 2019). I show how the common thread running through these authors, and largely summarized by Martin (2009), are in fact observations of this basic model of the management of uncertainty. I discuss how each of these authors relates to the Martin dimensions as well as to mine. I also discuss how the foundational properties of the deep Bayesian model connect to measures of political beliefs. I show how a more abstract model in a diverse environment describes a more liberal mode of political thought, while using a weak abstract model in a homogeneous environment describes a more conservative mode of thought. I will discuss the implications of these connections from a public policy point of view in Section 9.

Finally, I will show how certain real social networks, when mapped to the parameter space under discussion, give indications as to the political biases of the institutions who provide and tailor them. I close with a short anecdote in which I describe David Heise’s take on the (what was then) the two dimensional space.

My objective with this paper is to reveal a simple relationship between the way that uncertainty is handled in a broad class of two-level Bayesian models and the definitions of freedom, equality, and political belief that are most commonly used in the literature. I show how there are only certain configurations of such models that are functional, in the sense that the model’s posterior will cease to be informative should it not be configured according to a basic principle. Thus, I define a “manifold” or “subset” of such functional models, and I relate the models on the manifold according to the relative freedoms and equalities they (or their self-defined state) enacts. My hypothesis is that, since this this set of models has connections to both current models of neurophysiology and neural functioning in the human brain (Thagard and Schröder, 2014; Smith et al., 2018), and to theoretical models of reasoning based in neuroscientific evidence (Friston, 2010), this relationship represents a theoretical proposal that explains a variety of different sociological and political theories.

2 The Management of Uncertainty

Agents in an uncertain world are faced with a significant modeling challenge. Agents in a social world are even further challenged. If the numbers of interactants grow too large, the agent will meet its resource bound, and be unable to process any further interactants unless it drops some. This is well known to happen, but the number of agents that a human normally interacts with, say 150 (Dunbar, 1992),⁶ is far above what would be reasonably entertained in a computational machine the size of the human brain. For example, at only 15 other agents (the size of the sub-groups considered in (Moreno, 1934)), there are over 200 dyadic relationships to be modeled by each agent. Agents wishing to model triads are dealing with further complexity (a “third order” belief (Ridgeway, 2019)). Even restricting each agent to not model the contacts with anyone but herself, these 15 models will rapidly become very complex because of the missed links (correlations) between them. For example, if A is friends with B, and B is friends with C, then unless A models B’s relationship with C, there may be puzzling behaviours of B for A that are not accounted for by A’s relationship with B. These puzzling behaviours would require further modeling structure for A, but would not necessarily take into account known properties of C, and so may be impoverished.

Given these computational constraints, humans are likely to be able to only handle one, two or even up to five people at the same time (simultaneously in working memory, for example), and so these larger groups of 150 must be the result of some other process. By creating a hierarchical model, one can start to see a resolution to this impasse. Each agent can maintain a fairly crude, approximate, and general model of *all* others in its group, and use this as a rough guide to predict what everyone (including itself) will do. This rough guide is thought to be primarily computed in the default mode network, and will give indications (somatic markers) about what to do in what situations (Barrett and Satpute, 2013; Damasio, 1994). It is thought to be largely controlled by more “emotional” centers in the brain. This rough guide is a measure of the strength of alignment between agents.

2.1 Alignment

The basic idea of an agent who is aligned with its environment (its “econiche” in fact (Bruineberg and Rietveld, 2019)) is that the agent’s predictions (including of its own actions) must match its observations. When this happens the agent can predict the future is therefore more successful. A failure to predict future observations makes an agent vulnerable. Agents with better matching models avoid surprise and are better survivors.⁷ Agents with models that don’t match end up in a state of greater psychological entropy (Hirsh et al., 2012), which is known to be related to a state of increased anxiety, which puts an agent into a more exploratory (solution-seeking) mode.

Imagine an agent in a very simple world with two states, x_0 and x_1 . One way to picture this is the agent is a (not necessarily fair) coin. I will use a random variable X to describe the

⁶This number was implemented by Dunbar as the simple heuristic count of the number of people with whom you “regularly look into the whites of their eyes” (Dunbar, 2015).

⁷At this point there are no restrictions on this model whatsoever and the arguments presented still hold. Although I specialize to a Bayesian model, this is not a theoretical requirement from first principles, and is thus an arbitrary (albeit convenient, certainly for exposition) choice.

states of this agent’s world,⁸ thus, $X = x_0$ or $X = x_1$.⁹ The agent has some actions $a_0, a_1, \dots$ available to it (say “flip the coin”), and its task is to figure out what the world is most likely to be like if he performs some particular action a_i . That is, he wishes to compute $P(X'|X, a_i)$, the probability distribution over the state variable after the action is taken, X' , conditioned on the current state variable, X and the action taken by the agent. For our simple binary case, this distribution is a binomial, and so can be described by a single number $\hat{p} \in [0, 1]$. That is, the $P(X' = x_0|X, a) = \hat{p}$, while $P(X' = x_1|X, a) = (1 - \hat{p})$.¹⁰ I will use $\hat{p}$ to denote the model parameter, and p to denote the actual value of p in the world (which is unknown, but I will clarify based on context). Therefore if the agent can estimate p correctly (so $\hat{p} = p$), known as the “risk” in economics, then he can do as well as possible decision theoretically (as a reward-maximizing agent in an uncertain world).¹¹ This is because the uncertainty causing p to be not exactly 1 or 0 arises due to the actions of other agents or objects in the environment, but is *predictable* with the precision of p .

Certainly an agent can estimate p by interacting with the world and gathering statistics. In the case of the coin, the agent can flip it multiple times to estimate the probability that it will come up heads on the next flip. However, if the agent cannot estimate p properly, then he does *not* know p , but instead knows only $p + \delta p$ where δ is a noisy variable known as *ambiguity*. Ambiguity should be kept as a separate term to *uncertainty* (which is the entropy of P). The reason ambiguity is important as a separate concept, is that it is a factor determining when people rely more on social than individual learning, alongside problem difficulty and learning cost (Heinrich, 2016, p.64).

Why would an agent no longer be able to estimate p properly? For two main reasons. First, the complexity of the world the agent is dealing with has reached his internal capacity to model it (the size of the set of configurations X represents, 2 in our running coin example). Any further complexity beyond this must be merged in to the existing model, and will displace other concepts or make their predictions less precise. In our coin example, suppose a second coin is added with a (true) probability of landing on its x_o of q (and so $P(x_1)$ is $1 - q$), then our very impoverished two-state agent above will not be able to fully model the four resulting states, and will need to assume they are the same, likely settling on some estimate of the average of p and q , but the ambiguity of both will rise. These ideas are foreshadowed in Lewin’s *topological psychology*, where he states:

“It may not be possible to derive situation S_2 from the situation S_1 because situation S_1 which one chose was not sufficiently comprehensive and therefore certain facts

⁸Throughout, I will use capital letters to denote random variables, and small letters to denote values that those variables can take on, e.g. $y \in Y$. When writing probability distributions, I will use small letters, e.g. $P(y)$ is the $P(Y = y)$.

⁹A random variable is simply a label that can be assigned to one or more propositions. Thus “ X ” in the coin example should really be the statement “The face of the coin that is showing”, which can take on values “heads” (I have called this x_o) or “tails” (x_1).

¹⁰In the coin example, X (the preconditions) may be the null set, or $P(x'|x, a)$ may be uninformative as there are no predictive dynamics (one coin flip does not influence the next). A prior over X' would still be required, and can also be modeled with $P(X'|X, a_i)$. However, in general these dependencies do exist, and so the relationship is retained. In sociological parlance, *dynamics* is usually referred to as a *process*.

¹¹Rationality is defined in this paper as decision theoretic rationality: taking the action that results, in expectation, in the highest payoff according to some cost function, see Murphy (2021) for a recent treatment of Bayesian decision theory.

were neglected which in the interval influenced the system from the outside."(Lewin, 1936, p.69)

Lewin is simply pointing out that the psychological "life space" (here "state space") is open, not closed as are the systems studied in physics.

Consider a single grass-eating agent in a world with two states: grass and no-grass. As the distinction is important for survival, the agent learns to distinguish between the two by adapting sensory systems or by creating neural representations. Now suppose that sometimes, hidden in the grass, there is a tiger. Clearly distinguishing between grass and grass-tiger will also be evolutionarily important as it will allow the agent to reduce false positives (false alarm tigers, causing flight) and false negatives (missed tigers, causing damage). The complexity of the constructed model must match the complexity of the environment, as too complex models will overfit, leading to unnecessary tiger identifications (false positives), but too simple models will underfit, leading to missed tigers (false negatives). To further complicate the situation, the tiger may be hiding information (as in human conflict and war) by disguising itself as grass, or may be a part of a complex team of tigers (although this is more cheetah-like).

The second reason is the system itself (including both agent and world) may be inherently ambiguous (to the agent). Normally, this means there is some factor that the agent is not aware of, and may not be capable of being aware of without deep investigation. Such factors can be as simple as giving rise to unmodeled noise factors in observations, but may be due to other causes. The world is rarely *inherently* un-ambiguous (except at very small scales), but *appears* so due to the mismatch between the cognitive agent model and the complexity of the environment.

Ambiguity in the model or in the world must be handled by making one of two choices, and here we will run into a fundamental tradeoff between two types of uncertainty. First, I must introduce observations of X , and to do so let us consider our coin example again, but this time the coin flip result is not seen, but rather only heard. I ask you, dear Reader, to imagine a certain type of coin that makes a somewhat distinctly different sound based on which side of the coin it lands on. In this case, the observation ω_x will be the sound, while X is the actual denotation on the coin. In order to make this relationship, the agent needs a model of the sound given the coin, $P(\omega_x|x)$. Once so armed, it can handle ambiguity in two different ways. 1) use models that make weak predictions over X (so the $P(X'|x, a)$ is dispersed) and believe in evidence (so the posterior over X is largely guided by $P(\omega_x|x)$; or 2) models that make strong predictions and trust them over evidence, so the posterior is guided by $P(X'|x, a)$. These two choices are dependent. The agent cannot make strong predictions and believe in the evidence, because these have to "match up" in order to be detectable. That is, two probability distributions, when combined (multiplied) in a Bayesian fashion, result in a distribution centered at their average, which clearly in this case of a mismatch will result in an incorrect agent model of X . The agent can also not make weak predictions and discard the evidence, because it will not have a precise enough estimate of potential futures and thus will become vulnerable as its model will cease to be an accurate model of the world. Thus, we can see a basic tradeoff between these two elements of uncertainty (predictive power of the model and observational noise), with a decrease in one being necessarily followed by an increase in the other. These types of tradeoffs will form the basis of the connections I will make to the sociological structure of culture literature, and were also considered in the "flashlight" allegory presented in the introduction.

As noted above, the amount of ambiguity in an agent's model is directly related to its resource

bound and to the complexity of the environment. Environments with many other agents attempting to cooperate would be highly complex, and therefore contain much more ambiguity arising from the agent’s inability to model the large quantity of potential world configurations. An agent can either accept the increased ambiguity, and is then dealing with an “invalid” world (Kahneman and Klein, 2009), or can allocate modeling resources devoted to that situation (perhaps freeing up other resources of unused material). This latter option corresponds to gaining expertise in the domain, and so the agent would be labeled an “expert”, and the situation would become more “valid” from the perspective of his model (but maybe not from others in his environment). Thus, low (relative to the resource) complexity or high resource (relative to the complexity) environments are considered valid (decision theoretically (von Neumann and Morgenstern, 1953)), while high complexity or low resource environments are considered invalid (Kahneman and Klein, 2009). Finally, invalidity may increase simply due to cost. If some action costs a lot to do, such that an agent is prevented from doing it, then she cannot estimate risk, which implies ambiguity and thus invalidity. Humans are known to rely more on social learning in *invalid* environments defined in exactly this way, which means they rely more on others: “*when problems are difficult, situations are ambiguous, or individual learning is costly*” (Heinrich, 2016, p.64). Experiments show that as the environment becomes more complex, conformity effects are magnified, which indicates a stronger group, emotional, and social effect (Muthukrishna et al., 2016) (see also Section 9.2).

Validity can also be observed directly at the level of muscle memory. A classic example is walking on an escalator in the same direction as the escalator is moving. When one steps off the end of the escalator, the sensation is one of the predictive model of the environment not matching the evidence. Using any standard relative tradeoff between the two, the model predictions suddenly and catastrophically fail when the world doesn’t keep moving as quickly, and signals sent to your leg muscles are therefore incorrect (invalid) with respect to the actual world of the future, and so a stumble may occur, or at least a funny feeling of something being very wrong with the world. As (Russell, 1945, p.822) notes in a discussion of Dewey, such invalidity is a measure of surprise (see Section 2.1), and therefore truth itself can be measured as surprise.

Validity can also be associated with a lack of *positive freedom* (Anderson, 2017). I will review this in Section 6.1. Essentially, positive freedom entails a rich panoply of possible options. Lack of positive freedom is achieved by highly valid models, with the most valid of these being fully deterministic. At this point, there is no positive freedom, but maximum validity. Robinson Crusoe implements this, as his environment includes no others (prior to the arrival of Friday), and is assumed to be perfectly deterministic and discoverable such that a Cartesian science can uncover all of its rules. In the words of Anderson (2017, p.46), “there is nothing to do but eat coconuts.” At the other extreme, there is maximum positive freedom, at the expense of republican freedom (being able to do what you want) and negative freedom (having restrictions on your possible action set). I will discuss these dimensions in more detail in Section 6.1.

In population ecology research, the selection and adaptation of organizations leads to better isomorphism (matching) between the organizations and their environment. For example, in quickly changing environments, which are invalid from the perspective of the organization as a model as they lack expertise in the newly changed environment, this is more difficult to achieve, and so more generalist organizations will survive better. In more slowly changing or valid environments, more specialized organizations will fare better. This implies a match between the uncertainty in the denotative state (the domain of interest) for the organization and that of the environment.

More uncertain environments “call for” (e.g. non-aligned organizations are either selected out, or must adapt) organizations with more innovative, less formal teams (Hannan and Freeman, 1977). More certain environments call for more certain models: systems that “fix” procedures, ensuring the environment remains valid.¹²

The higher level model I have referred to is an approximation to the original model, and helps mitigate this complexity as it can modulate parameters of the lower level. So naturally as the complexity of the environment grows (for a fixed model it becomes more invalid) this higher level approximation will start to have a greater weight (simply because the dispersion at the lower level has grown). Nevertheless, the higher level must have some utility, or it must make up for the loss of precision at the lower level. The advantage comes through social signaling, which are direct indications from other agents about the higher level model. Essentially, agents are communicating to each other what the parameter space for the models they are attempting to share looks like.

Agents must nevertheless still sort out how to tradeoff these two levels of processing, and there is some leeway afforded by action that can be played with by each agent. Facing a tiger, an agent has the option to flee, and thereby to put itself in a position where the tiger problem has disappeared. Fleeing environmental context in a social setting can translate to arranging a social network to be more predictable by making it more homogeneous. In this paper, I hypothesise that there is a continuum of modes from connotative homogeneity coupled with denotative complexity to connotative heterogeneity coupled with denotative simplicity. I further hypothesise that this continuum of modes maps to a continuum of political beliefs. Other options are to fight, or to nihilate the tiger and simply stop caring about the damage the tiger could do (which would likely not work out well in this case).

To sum up, as the complexity of the environment increases, an agent with fixed computational resources runs into a bound that prevents it from modeling the added complexity. As this bound is reached, the predictions made by the denotative representation have increasing trouble matching the evidence from the world, resulting in more dispersed estimates of denotative state and an impoverished mechanism for predicting the future. Such an agent can handle this inability to predict the future by believing the predictions and ignoring the evidence (underfitting, leading to high bias, low variance behaviour), or by relaxing the predictions and believing the evidence (overfitting, leading to low bias, high variance behaviour). While the first agent will have difficulty adapting to change (but can see it coming), the second will have difficulty predicting change (but can adapt to it). I will denote the first type of agents (those that underfit) as “L” agents, and the second type (those that overfit) as “C” agents. An agent with a hierarchical model can do both, however, as it can create and track a lower-dimensional version of the denotative state that allows it to continue making predictions, albeit with reduced precision.

¹²Note that valid environments do not necessarily entail small state spaces, as there might be very low entropy probability distributions in the predictive function of the process. Thus, a large state space may be “visited” only in a small number of places, making the environment valid again. Such distributions will arise due to norms, customs, habits, enforcement, or any fixed rules to obey. As an example, consider that the number of possible ways of playing a chess game if disregarding the rules of movement (e.g. a Rook moves in straight lines) is vastly greater than the number of possible legal chess games. This is also known as the “graininess” of the environment in population ecology research (Hannan and Freeman, 1977).

2.2 Deep Bayesian Modeling

An agent can handle uncertainty with a straightforward symbolic model (i.e. a probability distribution as I used above), but cannot do so with ambiguity, and therefore must use a hierarchical model. That is, ambiguity can only be handled by also including a model of it. This new model can then be considered as a *parameter* of the underlying symbolic model, and each setting of this parameter (each value of the new model) corresponds to a particular probabilistic model of the world. Seen another way, a model of the ambiguity is a lower-dimensional version (or abstraction) of the state that allows the agent to continue making predictions, albeit with reduced precision, but across a wider range of (potentially) unknown states. In yet another view, given that an agent has control of a probabilistic model of the environment, and further has control over the parameters of this probabilistic model, then in fact the probabilistic model is a class or set of probabilistic models, with the breadth or variety of the set being dependent on the parameter settings.

The model of this breadth is normally considered a “higher” level in a Bayesian model because it can be considered as a generator of the underlying model. I will label this higher level model’s state space as Y (connotative). This state space, as I argue in Section 4, is continuous and 18 dimensional. However, I will throughout the remainder of this paper, consider it as a one-dimensional real function (on $[-\infty, \infty]$). Since Y is in fact a set of parameters, it will require a dynamic model (parameterized by $\hat{\alpha}$) and model connecting it to the denotative space X , which I have previously labeled the *somatic transform* (Hoey et al., 2020). The somatic transform is the model $P(x, y)$, and is parameterized by $\hat{\gamma}$. Finally, the validity of the environment as modeled by the agent will be parameterized by $\hat{\delta}$ (like the $\hat{p}$ of before, but now higher $\hat{\delta}$ means less uncertainty directly). Thus there are three parameters in this model: $\hat{\alpha}$ (connotative dynamics), $\hat{\gamma}$ (connotative-denotative connection) and $\hat{\delta}$ (validity).

Recall that each agent is making predictions about X based on a *set* of models defined by the parameter variable I have called Y . The relationship between this parameter value and $P(x)$ is that Y in fact defines a probability distribution over $P(x)$. In the two-state coin example this is a distribution of possible values of p . This distribution is arbitrary and simultaneously constrains both X and Y . One can immediately perceive the tradeoffs that must occur between the two levels. More precise estimates of $\hat{\delta}$ will yield more denotatively influenced models, and more precise estimates of $\hat{\alpha}$ will result in more connotatively influenced models. Finally, more precise estimates of $\hat{\gamma}$ will yield more harmonious models (where both X and Y “agree”).

The agent also must make observations of its environment through its sensors, which I will describe with a random variable Ω . This variable is actually a set of variables with two elements. The first, Ω_x , with values $\omega_x \in \Omega_x$, are observations of some component of the world, including shapes and objects in the environment, and other agents. The second, Ω_e , with values $\omega_e \in \Omega_e$, are observations of emotional signals from other agents, as read through their body language or their facial expressions. These emotional signals form a separate channel of input to an agent so defined.¹³ The reason this separate channel is plausible in real networks is because it allows an agent to “close the loop” and to receive direct feedback from other agents about the effects

¹³This sounds problematic, but recent evidence shows that visual inputs go through three pathways: one for object identification, one for object location and affordance, and a third for the dynamic aspects of social perception (Pitcher and Ungerleider, 2020). Thus, a signal separation based on content may provide a separation of observations in exactly the way I have assumed in *BayesAct*.

of its actions (through Y). These signals are termed “emotional” and ensure that the agents’ connotative spaces are directly linked. This “emotion language” (Turner, 2016), communicated in part through facial expressions and paralinguistics, allows agents to communicate what aspects of the denotative state are worth more fully exploring. As a result, the connotative state acts as a “flashlight” that illuminates the *same* part of the denotative state for all agents who share it, as in the allegory presented in the Introduction.

2.3 Parameter space and the Simplex

In order for an agent’s (approximate) model to predict its observations, it can go in three, mutually exclusive, directions. In valid environments, $\hat{\delta}$ is large, so p is strongly predictive of X (the distribution $P(x'|x, a)$ is undispersed in X'). Such a model can be learned by an expert that can (sufficiently) accurately predict the valid environment (or make it more valid). This may be because the environment is very simple, or because the agent has trained to become an expert in this environment (which may not be possible in highly complex environments). Agents operating in such a world will do better (decision theoretically) with fewer interactions with others so as not to overload their resource bound. In fact such agents will also receive some input from their emotional signaling system, but this will be largely interoceptive (e.g. need for food, sleep, sex), which will guide their actions if pressing enough.

In invalid environments, $\hat{\delta}$ is small, the agent is faced with ambiguity and has two choices: 1) use models that make weak predictions over Y (so the $P(y'|y, a)$ is dispersed, so $\hat{\alpha}$ is small) but strongly connect immediate evidence (relate Ω_x to Ω_e through $P(x, y)$, so $\hat{\gamma}$ is large); or 2) use models that make strong predictions in Y ($\hat{\alpha}$ is large) but only weakly connect immediate evidence (so $\hat{\gamma}$ is small). These tradeoffs occur for exactly the same reasons as the ones considered above relating denotative state to observations. The tension between these two dimensions ($\hat{\alpha}$ and $\hat{\gamma}$), for small values of $\hat{\delta}$, is a result of an attempt to match the behaviours of individual agents to a predictive model of emotional dynamics. If there are strong predictions of the meanings of events, then there can only be a relatively weak connection to the actual events themselves, otherwise the two may easily “miss” each other. Similarly, weak predictions of meanings and weak connections to denotative events results in a weak overall model that cannot achieve maximum reward. Thus, the inverse trade-off between the two dimensions reveals itself. I will later show how the two other relative tradeoffs (between $\hat{\delta}$ and $\hat{\alpha}$ and between $\hat{\delta}$ and $\hat{\gamma}$) are also governed by the same underlying uncertainty considerations, and are therefore mutually exclusive, making all three dimensions mutually exclusive.

In order to see how these three different approaches are mutually exclusive, I can arrange them as three dimensions that are orthogonal, and propose that social structures can develop on a *simplex* which connects a point on each axis (see Figure 2). This simplex may not be planar, and may not be scaled in the same way as shown here. In Section 4.3, I go over a simple example where it is more of a “dome” or “bowl” shape. Whatever scaling factors are utilized, the same tradeoffs will occur, and we proceed without loss of generality. I will describe each of the primary axes in the figure using the parameters $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$, as well as a short description of each as in the figure.

- High values on the *denotative dynamics precision* dimension indicate a very valid environment ($\hat{\delta}$ is high), which can be easily predicted by humans. Individualistic humans

will only work together in such an environment should they be controlled by some kind of authority. This is because they can agree to follow some common plan, but someone must arbitrarily define what the plan will be. At the extreme, this dimension calls for individualistic agents with no relational connections to others. However, a hierarchical social structure may be induced in which most agents only interact with a few others. As more agents are included as social partners, the ability to handle the denotative complexity breaks down, and so the denotative precision is forced to decrease. To return to my allegory of the flashlight boys, these boys have devoted their time to learn the exact most efficient location to shine their lights. As it is the exact most efficient location, they both compute the same one, and find each other instantly. In an abstract sense, these boys really need no flashlights, and in fact will be distracted by one. That is, they need no social contacts as these only decrease their needed resources for physics calculations. However, they do need to be sure the other boy is doing the same calculation somehow.

- High values on the *somatic transform precision* dimension indicate high levels of homophily ($\hat{\gamma}$ is high), but low levels of trust outside small cliques (e.g. families). Such agents will make weak predictions but expect consistent evidence. This dimension calls for a clique-like social structure. Connotatively, one boy has the flashlight on wide beam and the other on narrow. Larger scale cooperation also requires a monarch.
- High values on the *connotative dynamics precision* dimension indicate that agents make strong predictions and expect others to do the same, thus calling for trust and transitivity ($\hat{\alpha}$ is high). These agents make strong predictions but are tolerant of diversity, and calls for a multi-hub like social structure. Connotatively, one young fellow has the flashlight on narrow, and the other has it on wide.

The projections of each dimension upon the simplex yield a bisector line in the equivalent two-dimensional triangle, and these bisector lines (shown dashed in Figure 2) correspond to a variety of labels applied in the sociological literature. These are the type of axes used by Bales, Martin, Lawler, Durkheim, and Fukuyama. These labels are included in Figure 2, I will return to them in Section 7. They are shown in some detail in Figure 2, and then referenced and labeled in Figure 3.

While many authors have attempted to characterise social structures using two dimensions (e.g., (Tilly, 1998)), a few have worked in three dimensions and found more parsimony (e.g. (Bales, 1999)). I will review these works in Section 7, but here mention some authors have clearly described a simplex as a “Shepard plane” (White, 1992) or “ternary graph” (Martin, 2009), arguing that different positions on the plane refer to different “species” of “social molecule” (White, 1992). Further, the species can be sustained only through control over “all the uncertainties that impact it” (White, 1992, p.22), and that these different uncertainties or “embedding ratios” (White, 1992) are mutually exclusive in some sense, compensating for each other. I argue here that a “species” of “social molecule” can sustain itself because member agents can interact with other member agents at an equilibrium defined by how they manage uncertainty. Self-sustaining societies can therefore exist at any location on this simplex, and any society so defined will exhibit the same ability to manage uncertainty overall, but will be doing it in different ways. In the words of Haidt “there is more than one form of moral truth, or more than one valid framework for judging people or running a society.” (Haidt, 2001, p.130).

Figure 2: Simplex on $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ showing the various properties of the models at each apex (in blue) and along each of the three non-orthogonal bisecting dimensions which lie on the simplex (in red). The labels on the edges, as well as the lists in text boxes, describe the associations made between the settings of the agents and the behaviours of a social group of this type and the structures they form. The authors making these associations are abbreviated: JLM: (Martin, 2009) (with page numbers); F&E/Moreno (Feld and Elmore, 1982; Moreno, 1934). Those not abbreviated are Lawler et al. (2009); Durkheim (2014/1893); Bales (1999); Fukuyama (1995); Ridgeway (2019); White (1992).

- Apices:

$\hat{\alpha}$: coherence/transitivity: describes the ability to maintain a transitive relationship (i.e. to expect from someone and to give to someone else)

$\hat{\gamma}$: dependence/equality: describes the degree of equality in terms of meanings for an agent

$\hat{\delta}$: validity/rationality: describes individuality of agents

- Axes:

$\widehat{A\Gamma}$: is the side of the simplex from $\hat{\alpha}, 0, 0$ to $0, \hat{\gamma}, 0$, and describes “organic” solidarity (based on trust and shared meanings), labeled “invalid” in Figure 2.

$\widehat{\Gamma\Delta}$: is the side of the simplex from $0, \hat{\gamma}, 0$ to $0, 0, \hat{\delta}$, and describes low trust, medium similarity groups, labeled “intransitive” in Figure 2.

$\widehat{\Delta A}$: is the side of the simplex from $0, 0, \hat{\delta}$ to $\hat{\alpha}, 0, 0$, and describes high diversity and lack of dyads, labeled “unequal” in Figure 2.

A : $\hat{\gamma} - \hat{\delta}$ is the perpendicular to the $\widehat{\Gamma\Delta}$ axis, and forms a bisector from $\hat{\alpha}, 0, 0$ to the a point where $\hat{\alpha} = 0$ and $\hat{\gamma} = \hat{\delta}$. This is the differentiation axis and describes transitivity. In Bales this is the forward-backward axis.

D : $\hat{\alpha} - \hat{\gamma}$ is the perpendicular to the $\widehat{A\Gamma}$ axis, and forms a bisector from $\hat{\delta}, 0, 0$ to the a point where $\hat{\delta} = 0$ and $\hat{\alpha} = \hat{\gamma}$. this is the validity axis. In Bales it is the up-down axis. It also correspondds to Durkheim’s mechanical to organic solidarity.

G : $\hat{\alpha} - \hat{\delta}$ is the perpendicular to the $\widehat{\Delta A}$ axis, and forms a bisector from $0, \hat{\gamma}, 0$ to the a point where $\hat{\gamma} = 0$ and $\hat{\alpha} = \hat{\delta}$. This is the trust/mistrust, equality/inequality, dependence/independence axis. In Bales it is the positive-negative axis.

Figure 3: Dimensions of meaning in the simplex shown in Figure 2.

3 Social-Psychological Perspective

I start a section about social psychology with a distinctly group-based construct, politics. From the laws themselves to the people who make them, enforce them, and evaluate them, politics are based on some degree of balance between *freedom* and *equality*. As Rousseau (1762, p.96) points out, “*every system of law [has] two main objects, freedom and equality*”. However, there are many ways to balance these two elements. For example, conservatives value everyone’s freedom to act (e.g. to carry a gun), while liberals value everyone’s equality of action. One can easily see that one cannot be free and be equal in a society of others. If everyone is free, then everyone must be unequal. If everyone is equal, then no one is free. “*The liberty of some must depend on the restraint of others*” (Berlin, 2002, p.171). Therefore, I can write:

$$\text{freedom} = \frac{1}{\text{equality}} \quad (1)$$

You, the reader, may feel threatened by the simplicity of this equation. However, as I will argue in this paper, freedom is nothing but an estimate of the dispersion or concentration of posterior belief distributions in a hierarchical model. Different types of freedom express themselves at different levels. Such dispersions are termed *variance*, and the inverse (reciprocal) of variance is precision, and very precise distributions require people to be less diverse, more similar, more equal. Thus, the equation simply states that freedom and equality are measures of the same thing, in opposite directions.

Figure 4: Freedom-Equality axis showing free energy minimum.

I can therefore represent freedom and equality on a single axis, as shown in Figure 4. On this figure, a society could be set up anywhere along the line between freedom and equality. However, using free energy principles, I can show that there is a “sweet spot” at which the group functions most efficiently. This sweet spot, shown in Figure 4 with a ★, is a minimum of free energy for the group, and is defined by how uncertainty is managed in a group of agents. I will return to this in Section 3.1. However, we can proceed by simply remembering that there is a spot between freedom and equality at which a social group will naturally tend to settle, and this spot will be the group’s natural equilibrium on the freedom-equality axis. Another way to say this is that any group attempting to settle *away from* this sweet spot, will be beset by disagreement, or lack motivation, as I will show in Section 3.1. They will therefore be less efficient, and will tend to be dominated by groups who are at their sweet spot. Learning where an agent should situate its own, internal model of the world is something non-self-interested that an agent does, but it is something that benefits the group as a whole, and thus qualifies as a potential true form of group selection (Haidt, 2012).

In order to make this more concrete, I need to introduce some additional dimensions that will characterize freedom and equality in a way that will enable further analysis.

3.1 Uncertainty

The axis of freedom and equality actually corresponds to a fundamental trade-off in uncertainty.

Figure 5: Diversity at freedom and equality poles.

Consider the diversity of a population. We can represent diversity as a distribution over a range of human attributes as shown in Figure 5 above the word “freedom”. The wider this distribution, the more diverse the population. The one shown on the left in blue in Figure 5 has people spread across a wide range of attributes: it is very diverse, and has a lot of freedom.¹⁴ The distribution over the word “equality”, on the other hand, is much less diverse, so people are spread across a smaller number of attributes, everybody is very much the same, and there is much more equality.

The variance in the group has to translate into variance in each agent’s mind because of the good regulator theorem: every operational system has to be a model of its environment (Conant and Ashby, 1970). So, those operating in the society of diversity are more often going to run into diverse views of things, and therefore they will learn a more uncertain or “spread out” view of their society. They will therefore be more free to choose their own actions as there will be less constraint from the group level (as it is more spread out). Agents that live in the homogeneous society are going to have very precise distributions over the other agents in their group since everyone is similar. Actions are constrained, but equality and security are guaranteed. Security is guaranteed because, if everyone is exactly the same as you, then you can be very certain about things, you are in a state of pure equality, and you get pure security as a reward: you can predict exactly what is going to happen next. If everyone’s very different then you will be very uncertain about how people will act but you can be free to act in any way you want because it won’t stand out and people will know how to handle it.

Therefore, we have found two ways of doing things: precise and homogeneous, or uncertain and diverse. These methods correspond with our common notion of equality and freedom, respectively. Clearly we can’t have both, and each has its advantages. Being very certain about things is good because it allows decisions to be made. An agent’s certainty in something needs to be raised above a threshold for action. Being uncertain about things is also good because it allows an agent to be flexible towards changing situations and new and different people.

¹⁴I am deliberately mixing up the colors (red/blue) and directions (left/right) here, as will become apparent later on. I am doing this because the concepts of left/right and blue/red are not relevant once they are seen simply as reflections of uncertainty management in Section 3.3. Further, they arose for arbitrary reasons (such as a certain group just by chance was on the left side of the room).

However, these extreme values are difficult to sustain a social order over. The reason is that, in a state of perfect freedom, no cooperation is possible: there’s just too much diversity. Similarly, a state of perfect equality won’t succeed because everyone has to be identical, so that the system becomes very brittle to intrusions or exogenous changes.

In fact, there is a sweet spot, shown with a star in Figure 5, that trades these two off optimally in the sense that each agent is able to accurately model the group and the group is able to accurately model each agent. At this sweet spot, agent and group share a model and are best able to predict the future world.

This sweet spot in fact is the configuration of both agent and group such that the free energy (Friston, 2010) of the group is minimized, and it arises from an agent leveraging the second law of thermodynamics for its own benefit by amassing orderly states (information) at the expense of externalities (Jeffery et al., 2019). In some sense, the group has transferred as much energy as possible into maintaining a state of low entropy, a state of order, and not succumbing to chaos. The group and the individual are aligned in this case, and the heightened collective consciousness, regardless of how it is implemented, allows individuals to be more free to think, be and do (Patterson, 2014). I will return to this uncertainty trade-off when I discuss the *BayesAct* model. However, I can plot the free energy of the entire system vs. equality, and I find there is a sweet spot at which the system trades the two off optimally in a theoretical sense. This is the spot at which the group as a whole best ensures its survival.

In this context, one important component of the free energy is the alignment between the individual and the group. This can be broken down into two terms, one that computes the extent to which the group would have to change to fit the individual, and the other that computes the extent to which the individual would have to change to fit the group. As a function of the social uncertainty of the individual, the first term is smaller if the individual is more loosely defined, as it means the group has to shift less. The second term is smaller if the individual is more precisely defined, as it gives the individual the ability to “find” the group. Smaller free energy configurations are better because they ensure there is less “surprise” for the group.

Essentially it is easier for the group to “find” the individual (blue line in Figure 6) if he or she is loosely defined, or “flexible”, and it is easier for the individual to assimilate into the group (red line in Figure 6) if they are more precisely defined (assuming they are close to equilibrium). If they are far from equilibrium, then a wider distribution will be called for as well. If we also take into account the group *density*,¹⁵ then the group component becomes dominant and larger, it being harder for the whole group to shift towards the agent (blue dashed line in Figure 6), and the resulting free energy has a sweet spot that has shifted *upwards*, towards more freedom, shown in Figure 6 with $\circ$. Such a shift may also be caused by the *intensity* of the relationship to the group. Those relations formed in primary socialization, for example, may have much more intensity, and therefore a much bigger effect, than those formed in secondary socialization (Berger and Luckmann, 1967, p.152). Other components of the free energy are discussed in Section 4.3, which also contains a derivation of the curves shown in Figure 6. This matching problem is also modeled with the flashlight allegory from the Introduction.

¹⁵Density is similar to Durkheim’s *volume of collective consciousness*, and relates to the effective size of the group an agent interacts with. A much denser group involves many more interactions with many members of the group, and thus carries a greater weight (Berger and Luckmann, 1967, p.152).

Figure 6: Free energy as a function of freedom/equality. I plot the $\log(\text{free energy})$ for clarity only. Lower free energy is a preferable situation, but the energy created by the group trying to match the individual (blue line) is balanced by that created by the individual trying to match the group (red line) to give an overall free energy which has a “sweet spot” (minimum) at ★. The black line shows the free energy if these are traded off equally. The dashed lines show the situation in which the group is much less flexible because it is larger. Then, the “group” component of the free energy (individual stays fixed, group attempts to match, blue dashed line) is much higher, plays a bigger role in the resulting free energy (black dashed line), and shifts the “sweet spot” outwards towards more freedom at ○.

3.2 Mechanical and organic solidarity

In this section, I address the question of how to move a society from freedom to equality (or the other way). Suppose we are in a situation where everyone is free, and we want to move to a situation where all are equal. We have to introduce some dependence between all these free agents and this means introducing some form of what’s called solidarity. “Ideologies generate solidarity” (Berger and Luckmann, 1967, p.124), and thus we can distinguish at least two primary kinds of solidarity, called organic and mechanical (Durkheim, 2014/1893). These are also known as *relational* and *transactional* forms of commitment, respectively, in (Lawler et al., 2009).

In organic solidarity, we allow everyone in the social system to be diverse, but we require that everybody has to trust one another. So everybody can be very different but you have to trust that your neighbor is going to do the right thing. There is a role-specific distribution of knowledge, all institutions are differentiated, and a division of labour is possible (Berger and Luckmann, 1967). In mechanical solidarity, on the other hand, everyone is equal and everybody follows a single plan, for example a rational plan. No knowledge is role-specific, and all social actions are institutional (Berger and Luckmann, 1967). Mechanical solidarity is the dimension of homogeneity.

Francis Fukuyama has written extensively about the concept of *trust* (Fukuyama, 1995), which

he relates to Durkheim’s organic solidarity, and which he defines as the “*expectation that arises within a community of regular, honest, and cooperative behaviour, based on commonly shared norms, on the part of other members of that community.*”(Fukuyama, 1995, p.26). Fukuyama’s “norms” in this quote are later clarified as “arational social habits” (Fukuyama, 1995, p.41), which is more akin to non-rational heuristics. Much like Lawler’s relational commitments, trust can replace instrumental commitments (contracts) and make the economic process more efficient, and can “enforce” the solution of public goods social dilemmas like the free-rider problem (Ostrom, 2015). Further, Fukuyama explicitly points out that the “*particular character of those values is less important than the fact that they are shared.*”(Fukuyama, 1995, p.153) That is, a large group of humans can set up a system of values that allow trusting interactions, but exactly what those values are is not as important as the trust they engender.

Now the two solidarities may be orthogonal (I have not drawn them that way in Figure 7), but are definitely dependent in any agent or group. That is, you can’t have both of at the same time, although one can be changed independently of the other. The simplest way to see this is by considering differences in diversity. As diversity increases, mechanical solidarity is forced to decrease, and organic solidarity is therefore forced to increase in order to maintain the same overall degree of solidarity. Therefore, the two are traded off and must lie along a line that crosses the freedom-equality axis at the sweet spot (shown dashed in Figure 7). In fact anywhere along the line between these two axes is going to be a permissible social order with the same freedom-equality trade-off, the same sweet spot. We will see in the next section that this line is in fact a part of a three-dimensional surface. I will return to Durkheim in Section 7.1.

Figure 7: Freedom-Equality axis showing two solidarity axes and minimum free energy line (dashed). Mechanical and organic solidarity may be orthogonal, as in Figure 8

3.3 Three Freedoms

I now project the freedom-equality dimension from Figure 7 into a three-dimensional space, so that it appears as a dashed green line in Figure 8. Mechanical and organic solidarity are also shown (dashed red), but as I will discuss in Section 9, they can be projected on an axis of *illiberality* and one of *economic left/right* on the simplex (see Figure 20). Solidarity increases as you move from freedom to equality. Finally, a new type of solidarity is postulated, *artistic solidarity*, that runs from the origin and describes any social system that is disconnected from reality. Such a social system achieves solidarity solely through emotional meanings, but otherwise is anarchical. I note that artistic freedom is one dimension of experience that is normally experienced early in life, as differentiation in the mind on these subjective dimensions may be expressed as a fantasy in the real world, like a dream disclosed (Arnheim, 1954). Further,

Figure 8: Simplex on the three dimensions of freedom and equality. Also shown are poles of perfect freedom ■ and perfect equality •, The star is in the most central position possible for a social contract. Mechanical and organic solidarity are shown, along with a third type of solidarity, artistic (see text).

artistic solidarity may only be possible in a situation of economic surplus, which allow for some individuals freedom from objective challenges of subsistence (Berger and Luckmann, 1967, p.81). I will not discuss this further here, as I want too focus on the three dimensional projection onto the axes of freedom you can see in Figure 8.

The slightly rotated space defined by these three freedom axes, shown in Figure 8, is both more convenient to talk about but also more directly connected to the Bayesian model that I discuss in Section 4. Before I talk about these freedoms, I point out that regardless of how we talk about this space we have to end up on (or near to) this two-dimensional surface shown outlined in blue in Figure 8. This surface is a *simplex*, and represents the “sweet spot” of free energy.¹⁶ As with the single dimensional (linear) version in Section 3.2, there are numerous different ways of achieving the same freedom-equality trade-off. The trade-off shown in Figure 6 happens along all three freedom axes, meaning there is a true, ideal “sweet spot”, but that there are many “nearly-sweet” spots and local minima in which a society can easily become “trapped.”

Why is the space three-dimensional? Because there are three things at play: the individual, the group, and the connection between the individual and the group. Another way of saying this is the objective (the group) the subjective (the self) and the connective (membership in the group). Moral psychologists also refer to these as moral ethics of autonomy, community and divinity (Haidt, 2012). The representations in an agent’s brain or mind of the social context pervades reason and thought, and the way in which each agent in each context trades off the social and individual contexts will be defined by, and will define, the social order and thus reality: “*the relationship between the individual and the objective social world is like an ongoing balancing act*” (Berger and Luckmann, 1967, p.134).

¹⁶It is not actually a plane, but rather a spherical shape or bowl shape, see Section 4.3, but this planar approximation sweeps arbitrary scaling under the rug and gives us a useful analytical tool.

Let us consider now what these three freedoms mean (Anderson, 2017). Freedoms increase as you move down each axis towards the freedom pole at the origin (■) in Figure 8. Republican freedom means you are not subject to anyone’s unaccountable will, and is also known as *independence*. So as you move out along the republican equality axis, then everyone becomes equivalent and dependent. Normally this is done by making all dependent on a sovereign or a monarch, such that all independence is removed by subjugation to the monarch’s unaccountable will. However, a smart and honest monarch gives his subjects lots of opportunities (positive freedoms) and lets them have free choices (negative freedoms) but can intervene at any time to impose an arbitrary will to ensure everyone is steering in the same direction.

Positive freedom implies opportunity. As your group moves down on this axis, you get more and more opportunity. This happens because constraints are slackened at the group level, meaning uncertainty must be managed at levels lower down (individual) and higher up (at the corporate or government level). As you move up this axis, then opportunity becomes more constrained. Positive equality is a place where everyone is exactly acting in the same way and the world is predictable and *valid* (Kahneman and Klein, 2009). Therefore if you could maximize positive equality, then everyone would act according to a single plan. One such plan could be a rational plan. By defining what is good and what is bad, a rational decision maker can be used to set policy. The definition of what is good and what is bad is still needed, and as I will discuss further in Section 9, may be used by a despot for personal (perhaps only status-based) gain.

Negative freedom is defined by the freedom to choose your own actions, from whatever choices you are given. So moving towards negative equality means removing people’s abilities to choose their own actions. One way to do this is by defining affective identities, and then making more stringent requirements on how actions should be coherent with these identities. These culturally approved dynamics become institutionalized, and they remove negative freedom of individuals to act in whatever way their will directs them. Thus, an increase in emotional coherence between (seemingly self-imposed) actions and behaviours, in an emotionally stratified society, leads to a reduction in the space of actions under consideration, accompanied by a corresponding increase in negative equality in which actions are constrained by social prescriptions. A state of “world closedness,” extracted from a state of “world openness,” is a result (Berger and Luckmann, 1967, p.51). The classic imposition on negative freedom is private property. So I can wall off a piece of ground for myself, and I have increased my negative freedom on my property.¹⁷ However, I have decreased the negative freedom of 7 billion people (well, realistically probably only a few hundred co-citizens of my rural town), and therefore overall have increased the negative equality.

The ultimate republican equality is equality under some god. From the perspective of the infinitude of any supposed god, humans will look all equal. To create a functioning society then, a god may be replaced by a human deity, such as a monarch. Such a system implements perfect republican equality, or the perfect *lack of* republican freedom: one’s freedom may be curtailed at any time by the monarch (by e.g. cutting off your head). We may also be tempted to call this “oppression”, but this has a negative tone that is not necessary in a society defined this way. However, such a system also removes all *negative equality* (it implements perfect *negative freedom*), because everyone is free to do as they please, knowing that any trespasses (which are not necessarily clearly defined) will be punished by the monarch in an unpredictable way. That is, each person is free act as they please, and there are no inalienable rights, something I will return to in the next paragraph. Finally, such a system also removes all *positive equality*

¹⁷I still require the law, and an enforcement component of government to ensure this freedom is upheld.

(assuming a benevolent monarch), because many opportunities for growth may be opened up due to the simplicity of the ruling system, thus maximizing positive freedom.

3.4 Self

Before I continue our discussion of freedom, I want to discuss the related notion of *self*. These two notions are intertwined in a peculiar way (Berlin, 2002). From a certain perspective, laws made by the state are definitions of a generic, unified, social, self, which may be more or less well defined. Thus, the person making the laws is, in some sense, defining what it means to be a member of the group. If someone is found who does not conform to this definition, they are labeled an outcast and find it difficult to make progress by exploiting social structures because they are explicitly excluded. This can be viewed as a defense of the freedom of all individuals to act as they so choose. “Why a defense, since it restricts freedom?”, the astute reader may ask. Indeed, a defense, for in order to maintain a society in which all individuals are free to act as they choose, one must also institute a set of rights (a definition of self) that protects each individual from the imposition of any constraints on their actions by others. Once this set of rights is agreed upon, each individual has freedom to act as they choose, so long as they respect the rights of others. This establishes that increased inalienable rights implies decreased negative freedom *for the individual*, but this is compensated for by an increase in positive and republican freedoms.

However, a greedy autocratic state can easily shift this definition, making it more and more stringent with the argument that this actually increases overall freedom, although it is only increasing positive freedom, while decreasing negative freedom by defining the self more precisely. The autocratic state can further support its argument by noting that, in fact, “some [animals] are more equal than others” (Orwell, 1945), and in fact the increase in laws is only a reaction to the lack of virtue in the citizens, who must therefore feel guilt and conform to be virtuous. Those that cannot conform are labeled as enemies of the state or heretics. Such a state can then establish the one “true” citizen that all must aspire to. Marx’s writings show this ideal citizen to be clearly defined in the definition of the communist state. “*The Communists, therefore, are on the one hand, practically, the most advanced and resolute section of the working-class parties of every country, that section which pushes forward all others; on the other hand, theoretically, they have over the great mass of the proletariat the advantage of clearly understanding the line of march, the conditions, and the ultimate general results of the proletarian movement.*” (Marx and Engels, 1967 (1872, p.95). Unfortunately, this may have led to the development of autocratic communist states in which much suffering was inflicted in the name of negative freedom. The take-over by the “politburo” (however defined) is a shift across the simplex from the $\hat{\alpha}$ apex towards the $\hat{\gamma}$ apex, as republican freedom is given up for an increase in negative freedom. Communist and Fascist state definitions exist at this far corner (high $\hat{\gamma}$, low $\hat{\alpha}$ and $\hat{\delta}$).

Consider the role that the environment plays in the definition of the strategies to be carried out by these ideal citizens. The environment does two things: first, it introduces genetic diversity; and second, it introduces uncertainty. In fact, in a social world, genetic diversity also fundamentally introduces uncertainty. This uncertainty conflicts at a very basic level with the definition of the “advanced and resolute” citizen with “clear understanding” that Marx sought. Since, at this perfectly rational ideal, all strategies must unify into one categorical imperative, then, as Berlin notes: “*a rational solution of one problem cannot collide with the equally true solution*

of another, for two truths cannot logically be incompatible." (Berlin, 2002, p.192). Thus, should this one "true" citizen be defined, then all would be acting according to the same strategy, and so be perfectly harmonious. "*Universal justice which springs from reason alone*" Rousseau (1762, p.80) would prevail, which, according to Rousseau, came from God, although immediately admitted as unusable by humans as such, but might be approximated to some degree by the state (Rousseau's "sovereign").

One would label such an environment as *valid* (Kahneman and Klein, 2009). Although a person in such a state would have negative freedom (he is free to choose what he does, within boundaries set by the rights of others) and republican freedom (no one can impose an arbitrary will on him), but no positive freedom as his strategy is pre-defined and his life may be explored only according to that one policy, and so opportunities are decreased. At this apex with no positive freedom, positive equality is maximized, and all people have the exact same path charted out for them through a definition of what is a "true citizen". This of course leaves the question open as to who decides what the "true" citizen looks like, and in this is a communist state's downfall. Berlin notes "*Freedom is not freedom to do what is irrational, or stupid, or wrong.*" (Berlin, 2002, p.194), thus implicitly requiring a definition of irrational, stupid and wrong. This definition, of what Friedman called "madmen and children" Friedman (1962, p.33), is left open, but may appeal to politically minded persons with autocratic tendencies. Notice that the restoration of republican freedom by such a regime is only within the rigid definition of a Marxist state, and thus, although present, has limited impact on any one individual, although it is true of the society.

Thus, an argument can be made that it is the definition of the self that is at the root of freedom, and therefore, if you are prepared to indulge me by accepting my three dimensions of freedom, which I hope to convince you of in the following pages, you should be prepared to accept an isomorphic three dimensions of self. What are these three dimensions of self? In this paper, I attempt to show that they are in fact different ways of managing uncertainty in the relationship between the internal, "subjective" world of the mind, and the external "objective" world of the environment, however defined.¹⁸ Uncertainty is important at a very basic level, because its increase corresponds with a decrease in survival chances.¹⁹ Uncertainty arises from many sources, some of them beyond human control (such as an earthquake), but many due to the social environment of the person. Uncertainty also arises from internal sources, such as ruminations on one's own death (van den Bos and Miedema, 2000), but also sickness, hormonal changes, etc. Such uncertainty may also be caused by the actions of others, whom you cannot control and who are largely unpredictable. A single person, living alone on an unreachable island, may control his environment in such a way as to perfectly reduce this uncertainty (Defoe, 1719). He thus controls all uncertainty in the objective world, creating a situation of perfect negative and republican freedom (he can do as he chooses, subject to the vagaries of nature, over which neither him nor any other man has any control), but lacks all positive freedom, because "*there is nothing to do but eat cononuts*" Anderson (2017, p.46). Living in a society, a person must accept infringements on his negative freedom from other people who also lay claim to these same freedoms. Thus, I cannot decide to threaten my neighbour, because I must grant him the inalienable right not

¹⁸I put "subjective" and "objective" in quotes because they refer to one perspective on this only (the other being the inverse, but it makes no difference to the argument).

¹⁹This is at the level of eating or being eaten. Uncertainty has benefits because it can open up new opportunities, but only if those are truly opportunities, and not damaging to the individual.

to be threatened that I enjoy.²⁰ My negative freedom is thus immediately infringed upon, but I gain the positive freedom that accompanies interactions with other intelligent agents. What I am really doing when I make this change is relaxing my control over the environment, allowing others to do unpredictable things, by managing the uncertainty at a subjective level, in the mind. That is, I must become more convinced of who I am and what my role is in relation to you, essentially the relationship between my freedom and your rights (and vice-versa). I loosen my grip on the world by tightening my grip on the self. I loosen my grip on the objective by tightening my grip on the subjective. I seek to show that this corresponds to an increased reliance on the *self*, and therefore an increased set of rights. Finally, a third way of managing the uncertainty is possible, and this is in the *relation* between the objective and the subjective. So long as I manage to keep the objective and the subjective coherent, then I am not assuaged by doubts and uncertainty about either taken apart, for they are always together.

Note that in Figure 8, there may be an arbitrary, unavoidable, scaling on each dimension. However, regardless of any three multiplicative scale factors, one for each dimension, one can always map it back to the equilateral simplex depicted in the the figure. Thus, one could easily “stretch” this space so that the three dashed red axes of rationality, economic left and illiberality. This also implies that the theory I am presenting is scale-free because of this arbitrary scale factor that would need to be determined empirically. It is therefore less relevant where on the simplex any particular social system exists, rather what matters are changes across this simplex.

3.5 Freedom and Equality

The three freedoms and equalities are thus simply measures of where in the continuum from objective to subjective is the uncertainty managed within each individual. Groups of people, as I will show, will do better if they are consistent with each other in this management strategy. They may be represented by a state, which would be able to carry out their wishes and set just the right amount of laws and impose just the right amount of control, such that the people in the group feel they are being treated equally *and* have freedom that they value according to how they manage uncertainty in their own minds. Now consider what happens when multiple groups attempt to do this. This would seem to work, so long as the groups stay apart from one another. Once they start to interact (e.g. trade with each other, having romantic relations with each other), then all bets are off because the two states disagree, and thus become two *parties* who together define the state. These two parties must now come to some equilibrium, and many societies through history have attempted to create mechanisms to facilitate the finding of this equilibrium.

Once this theoretical correspondence is made between the management of uncertainty and the ideal states that result, then I can use the models so defined as artificial intelligence that is guided by a social sense of “self”. The definition of this model is in Section 4. First, however, I will show how the management of uncertainty in any two-level Bayesian model defines an implicit three dimensional space of uncertainty: one dimension for each level in the hierarchy and one dimension for the relation between them. “Why Bayesian model?” my astute reader asks. Because of a growing body of evidence that the operation of the brain may be something

²⁰Even if my neighbour does not have that right, he may still have the right not to care about being threatened, thus freeing himself from my removal of his negative freedom. If he takes this further, he may become a ascetic monk and remove his own life, granting him perfect freedom.

similar to a hierarchical Bayesian reasoning engine (Friston, 2010). Nevertheless, I can still indulge the sketpic reader my waving my hands about and claiming that these same principles, regardless of *how* the uncertainty is managed, still apply to any sufficiently complex system: the uncertainty must be dealt with at some point in the model. The Bayesian framework gives us an elegant theoretical basis on which to explain how such a reasoning engine would solve the problem.²¹

3.6 Security and Rights

Two other important concepts are security and rights. I will consider each in turn. Security is closely related to equality. If all are exactly equal, there is perfect security. The opposite of security is insecurity, a synonym of which is uncertainty. At the origin of the simplex in Figure 2, uncertainty is maximal in all three dimensions, so that insecurity, freedom, inequality, individuality, are maximized, with death being the ultimate freedom, while security, servitude, commonality, and equality are minimized. At the other extreme, uncertainty is minimal (high $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$), so that security, servitude, commonality, and equality are maximized, with Smurf village (without Gargamel) being the ultimate security.

Generally speaking, negative freedom corresponds to a lack of rights for the group. As one moves the group out the $\hat{\alpha}$ axis, negative freedom drops, and these are societies with more individual, inalienable, rights. For example, private property is a right to a certain piece of the land. The owner of this land has negative freedom on his land at the individual level, but the population as a whole has lost negative freedom (because N-1 others have their actions restricted to *not* be on this individual’s piece of land. Capital punishment cannot be administered if life is an inalienable right. However, negative freedom can also mean the freedom to give up one’s freedom (freedom to be a slave), and therefore at the individual level, negative freedom includes alienable rights (the individual can give them away). Therefore, one must keep the individual model separate from that of the group in any analysis, see Section 2.1, and high $\hat{\alpha}$ corresponds to inalienable rights.

3.7 An example

In 2020, a pandemic caused many countries to “shut down”, that is, to close all businesses, massively restricting the positive freedom of the citizens, while leaving them their negative freedoms (mask wearing not included). That is, everyone was free to go wherever they wanted, its just that everything was closed. This method was selected, I believe, because it is easy to police as there are few stores compared to people. The alternative methods are restrictions on the other two freedoms. First, suppose all businesses are allowed to stay open, so positive freedom is upheld. However, individuals are restricted to go only on certain days. I’m a “Friday person” or a “Fridayer” as they came to be known. I can only go shopping on Fridays. Everything is open on Mondays, but I am not allowed to go. This is more difficult to police as everyone shopping needs to be checked. However, if I really embrace my new identity, I will not longer even

²¹Note that I am not necessarily *not* being reductionist here. I still believe, at its very core, that all reasoning in the brain takes place by pulses within and even perhaps between neurons. Nevertheless, I also believe that social pressures allowed the evolution of fairly complex structures in networks of neurons that exhibit Bayesian-like behaviours in humans. Therefore, by studying the behaviours, we seek to provide a reference, goal, or method of evaluation of more reductionist network models that are more difficult to interpret.

want to go shopping on a Monday. What kind of a Fridayer would go shopping on a Monday? Outrageous! So I adopt a new identity, and I don't care about shopping on Mondays anymore, so I recover my lost negative freedom. This seems preferable, but recall that the group as a whole has lost negative freedom. Note also that although this is harder to police, once the moral norms involved with the identity of "Fridayer" (that they would never be seen shopping on a Monday!) has been established, people's commitment to it becomes relational, stronger, and longer-lasting, therefore requiring less policing.

The final choice is to restrict republican freedom. In this case, positive and negative freedoms are upheld, everything stays open, all people are free to go when they want. However, the supreme ruler employs posses that roam the streets and arrest anyone who spreads the virus. The definition of "spread" is loose, implying control of the virus spread through fear in the people of arbitrary retribution by the leader. This method (like the others) will work better with some groups, and less well with others.

4 Affect Control Theory and *BayesAct*

BayesAct is a two-level model of human intelligent and affective reasoning (individual and social) that explicitly represents the three dimensions of uncertainty in a simple and measurable way by leveraging the machinery of affect control theory (ACT) (Heise, 2007; MacKinnon, 1994; Heise, 2010; Hoey et al., 2016; Schröder et al., 2016; Asghar and Hoey, 2015; Hoey and Schröder, 2015). ACT is a model of emotional coherence based in language that was founded on the control principle of Powers (1973), which states something very reminiscent of the free energy principle: that people try to minimize incongruencies by controlling their perceptions. He transposed this to the sentiment space of Osgood et al. (1957), imposed a denotative structure from symbolic interactionism (Mead, 1934), and added affective dynamics (Gollob, 1974). ACT is a computational model that has been used to predict classes of human behaviour in a variety of settings (MacKinnon and Heise, 2010). ACT maintains a deterministic and static denotative model as an actor-behaviour-object-setting state (e.g. *doctor advises patient in clinic*, and an associated deterministic but dynamic connotative model: a dynamical system in Osgood's three-dimensional space of affective meaning: evaluation, potency and activity. This dynamical system represents *values*, or *evaluative knowledge*, which can be contrasted with *declarative* and *procedural* knowledge that are represented in the denotative model.

ACT proposes a fundamental link between symbolic, denotative, representations in the mind of the social environment and the continuous, connotative, representations of the sentiments or feelings associated with those denotative representations. For example, when one perceives a person in a white coat in a hospital, a denotative impression is formed of this person that is represented with a symbol (*doctor*). This symbol has an associated *fundamental sentiment* in a three-dimensional affective "EPA" space of evaluation (E: good/bad), power (P: strong/weak) and activity (A: active/inactive). The word *doctor*, for example, usually evoke feelings of goodness, strength, and modest activity. EPA space has been found through decades of research to be a cross-culturally normative representation of meaning (Osgood, 1969). The link between denotative and connotative in ACT is empirically determined through population surveys using semantic differential scales.

ACT includes a control mechanism, the *affect control principle*, which is defined by the difference

between fundamental (out-of-context) sentiments and the in-context *transient impressions* that are formed when events co-occur (e.g. an agent doing something to another agent). ACT’s control mechanism seeks alignment (agreement) between these two sets of sentiments. This optimization reveals the connotative meanings of the optimal action for an agent to take next. These connotative meanings can be communicated through signals of emotion, which are coded indications of the differentials between in- and out- of context sentiments. These differentials give other agents the ability to optimise the overall alignment.

BayesAct (Hoey et al., 2016; Schröder et al., 2016) generalises ACT by explicitly representing the distribution over sentiment in a two-level partially observable Markov decision process (POMDP) (Åström, 1965; Boutilier et al., 1999), a model extensively used in operations research (Puterman, 1994). *BayesAct* models individual differences as variance in sentiments, and includes a denotative state that models all elements of an interaction not relegated to the affect control system. The somatic transform connects the connotative with the denotative space, and so the optimization over actions happens simultaneously in the two spaces. As the connotative and denotative spaces have different dynamics (predictor functions), it rests on the *relative* strengths of the predictions in each, coupled with the *relative* strength of the connection to determine which of these systems will carry more “weight” in the determination of action. Once this linking occurs, then the connotative state and dynamics encodes a *social contract* with the social group in which the agent finds itself. This is the primary subject of my discussion in the last section, and will be more completely formalized in the next, where I present the *BayesAct* model in an abstract form, and refer you, dear Reader, to the prior work in (Hoey et al., 2016, 2020) for further information and derivations. This abstract form is meant to reveal only the necessary tradeoffs to make the connections to cultural sociology that I seek. However, in the full *BayesAct* model, the connotative and denotative spaces are factored into different pieces, with potentially complex relationships between them, both statically and dynamically. The full model factors Y into fundamental sentiments (9 state variables) and transient impressions (another 9 state variables), and models the relationship between them with another somatic transform that represents the *affect control principle* (see above).

A POMDP instantiates a *temporal frame* or *structural representation* (Russell and Norvig, 2010). Frames, as schemas, are a classic structure used in early artificial intelligence (AI), Knowledge Discovery and Data Mining (KDD), and Information Retrieval (IR) research that assigns a label and interpretation to each object, fact, relation and event that constitute a particular situation. Such structures are typically propositional and discrete-valued to enable ease of use in a computer program. For example, we might label the positions of pieces on a chess board, or predictions about how a game will turn out given a sequence of moves, or the bids in a negotiation. Policies (strategies) can be computed using decision theoretic principles, in which expected values are computed and compared for different strategies of action. These expectations combine probabilistic estimates of the likely outcomes with preferences over those outcomes.

The inclusion of ACT means the model must be augmented with labels for identities and behaviours corresponding to ACT’s denotative model, but with added noise modeling. These labels can then be interpreted as distributions in a sentiment space using a measured dictionary. This sentiment space thus complements the *denotative* state I have been describing so far, with a *connotative* state (which in fact is 18 dimensional). The model is fundamentally based on the symbolic interactionist idea that symbols (language) provide order for “the subjective apprehension of biographical experience” (Berger and Luckmann, 1967, p.97). Symbols are then reified

elements of exactly these same subjective apprehensions. Thus, learning and being become one single experience. The combination of symbolic and affective interpretations is what enables *generalization*: once the symbol “doctor” is assigned to someone, expectations for her behaviour become defined as generally as possible with respect to her occupation. That is, I expect her to do something good and powerful, but I’m open to a range of actual objective actions that could be in play in the current situation. For example, if the current situation is a court-room, I still expect her to do something good, such as testify honestly, and powerful, such as speak authoritatively.

Frames form the foundation of much knowledge representation work in AI, but have been efficiently implemented using Bayesian networks (BNs), which can be used to compute a distribution over all possible worlds modeled by a particular frame (Pearl, 1988). This probabilistic model then rests on the structural ontology and temporal logics that are proposed in the frame. Bayesian decision networks generalize the *goals* in frames as preference functions that rank all possible outcomes using a numeric scale, e.g., a utility function (von Neumann and Morgenstern, 1953). *BayesAct* complements this denotative model (the variance of which is the *invalidity*), with the ACT-based connotative model (the variance of which is the *coherence*), and a model of the relationship between them, the *somatic transform* (the variance of which is the *dependence*) (Hoey et al., 2020). For example, in a government policy decision, the objective facts may include the amount of money spent or saved, and long-term estimates from potentially complex predictive models, and the utility is financial. The denotative temporal dynamics may describe immediate and longer-term effects, enabled by adding more latent state, and allow for the construction of a policy that optimises over some definition of utility based on the same features. For example, a government may consider whether to increase the size of primary school classes. This will save tax and not unduly inconvenience teachers or students. The utility may be the overall savings in tax that can be passed on to the “man in the street”. The denotative temporal dynamics may also encode *norms of behaviour* that indicate the normative choices to make for any given identity-behaviour combination (e.g. a “citizen” should not “free ride” on other “citizens”, or a more extreme example, a “son” should not “make love to” his “mother”, which incidentally is emotionally coherent, but denotatively taboo).

The connotative dynamics are ACT-based, and will encode the relative freedom tradeoffs for whatever group they are applied to. That is, for some particular configuration of the denotative state, including a definition of identities, a connotative distribution results that may be used to compute how emotionally coherent various behaviours are. This connotative coherence is one of “feeling” or “intuition”, which may override any norms. I will call such coherence “prescriptive” rather than “normative”. A striking example is a trolley problem, in which it is logical to throw the switch on a runaway trolley so that it kills only one person instead of five, a strong connotative prescription against “killing someone” may take over for many people and prevent this logical strategy. Another example is an ultimatum game, in which one player is given \$10, and can give any amount he wishes to the other player. While logically the proper amount to give is \$0 (or 1¢ if the game is repeated and the other player has a choice not to play), most humans will fork over about 20%-40% of the amount they are given.

This connotative distribution, and the resulting affective predictions, constitute an *institution*, defined as a “*reciprocal typification of habitualized actions by types of actors*” (Berger and Luckmann, 1967, p.54). The coherence of this typification is then traded-off in *BayesAct* against the validity of the denotative state, according to the dependence encoded in the somatic transform.

Each group being decided for may have a different interpretation of the situation. It is this interpretation that then determines which options are considered, as they will be those that are emotionally most coherent. Thus, a liberal-minded person will view the classroom size increase as something that will hurt children, causing a strong incoherence emotionally. A conservative-minded person will focus instead on fairness and freedom, and will therefore not perceive the child as being “hurt” by the classroom size increase. In this case, the very high incoherence created for the liberal (who, as you will recall, requires highly coherent social relationships) may blind her to the advantages of the increased classroom sizes. A similar effect is noted in work on self-affirmation theory (Sherman and Cohen, 2006), in which uncertainty about the self (incoherence) leads to less objective evaluation of options, and more subjective use of stereotypes and emotional biases. The conservative on the other hand, focusses more on the sentiments of things in the moment, and will not perceive the child as being hurt by a classroom size increase (even the opposite: more friends!), but will see their neighbour being hurt by not being able to afford Christmas presents. Thus, being more self-affirmed, they may be able to more objectively evaluate the two options and rationally decide that a classroom size increase will have a net positive benefit.

Self-affirmation is not an individualistic enterprise. It is, in fact, a process of “fitting in” with a group, of the group and the individual becoming “mirrors” of each other (Mead, 1934). Well fit groups form a community of practice (Wenger, 1998), and have routines and identities that members can lock on to. These routines then increase self-affirmation, and allow for a division of labour within the group, and “opens the way for innovation” (Berger and Luckmann, 1967, p57). Similar effects have been noted across a range of studies since then (Muthukrishna and Heinrich, 2016), although it is possible that over-doing it may also be detrimental (Lazer and Friedman, 2007). In *BayesAct*, a move towards organic solidarity (increased coherence at the expense of decreased dependence) also allows for greater positive freedom, and more flexibility in coming up with innovative, exploratory solutions. An agent with a well-fit model will be self-affirmed, and will have more denotative freedom leading to innovation perhaps. The poorly fit agent will be in a state of uncertainty about the self, and will rely more on denotative meanings, which may contain stereotypes. The institutions created in organic solidarity form a protective cocoon (Giddens, 1991), or a “shield against terror” (Berger and Luckmann, 1967). The ultimate such institution are religions, which form a “sacred canopy” (Facciani, 2020).

4.1 Deep Bayesian Models

Figure 9 shows a graphical model representation of *BayesAct*.²² This graphical model shows random variables as circles, decision variables (under the control of the agent) as squares.²³ Undirected (directed) links give probabilistic (conditional) dependencies between variables. Thus, $\textcircled{X} - \textcircled{Y}$ represents $P(X, Y)$, while $\textcircled{X} \rightarrow \textcircled{Y}$ represents $P(Y|X)$. The product of all the functions so defined in the model is the joint probability distribution over all the variables. The semantics of this type of graphical model is that it leverages the conditional independence in the graphical model in order to compute this potentially very complex joint distribution.

²²Such models are a formal representation which is useful for human prior specification of the domain, the standard reference is Pearl (1988), although treatments can be found in any artificial intelligence textbook (Russell and Norvig, 2010), the most comprehensive being (Koller and Friedman, 2009). For MDP-specific material, see (Puterman, 1994). For POMDP-specific material refer to (Boutilier et al., 1999).

²³Rewards or utility functions are usually shown as diamonds, but are not shown in Figure 9 to reduce clutter.

Figure 9: Bayesian decision network for *BayesAct* at a high level of abstraction showing denotative X and connotative Y , observations Ω_x , emotions Ω_e , and actions both denotative A , and connotative F_b . Directed links imply (but do not require) causality in a Bayesian network. Two somatic potentials, modeled as undirected links (see text), link state and action, respectively. Primed variables are post-event, and the network is dynamically unrolled through time (that is, it repeats over and over again, one such “slice” for each discrete time step. The reward function is not shown, but would connect to X only.

The graphical model is defined for a discrete temporal sequence, with one set of variables for each time step. Two time slices are shown in Figure 9, with primed variables indicating those that are in existence after an action is taken, but in reality the network is “unrolled” for as many time steps as required, and the dynamics is considered stationary (does not change over time).²⁴ A Bayesian decision network, along with associated probabilities and utility functions, can be used to answer queries of the form $P(X_t | \Omega_0, \Omega_1, \dots, \Omega_t)$, the posterior distribution over the denotative state at time t given observations up to that time of $\Omega_0, \Omega_1, \dots, \Omega_t$. A number of algorithms exist for computing these inferences, and these can be computed efficiently and recursively for structured models such as a POMDP by repeatedly taking products and sums of conditional probability and utility functions. Optimal decisions can also be computed using similar algorithms, where maximization over the action variables takes place of summations for random variables. However, these computations are significantly more difficult than those for inference, and normally require approximations such as stochastic simulation (Monte-Carlo tree search), see (Hoey et al., 2016; Asghar and Hoey, 2015; Silver and Veness, 2010). Such models also come with a formal learning theory in which model parameters may be estimated based on statistics of the associated observations. In fact, it is this learning that turns the *BayesAct* model into a Bayesian reinforcement learning system (one in which the parameters of the model, aka the parameters of the different conditional probability distributions, are also included as variables in the Bayesian network). Since the connotative state Y “guides” the search through denotative

²⁴Although by deepening the network, one can account for changes in, and learning of, these dynamics, leading to a hierarchical POMDP that can be treated as a Bayesian non-parametric model - the infinite POMDP (Doshi-Velez, 2009) - and Bayesian reinforcement learning (Duff, 2002).

space, only those “illuminated” regions will be learned.

The connotative state, $\mathbf{Y}$, includes both fundamental sentiments (EPA ratings for words) and transient impressions (EPA ratings for words in an actor-behaviour-object event), as well as somatic transform links parameterized by $\hat{\alpha}$ between them. These links encode the affect control principle, which states that in- and out- of context sentiments should be as similar as possible. The connotative behaviour variable, $\mathbf{F}_b$, has been singled out in Figure 9, but really is part of $\mathbf{Y}$. The denotative state, $\mathbf{X}$, includes a component that describes the problem space or domain of the interaction (e.g. chess board pieces if the interaction is a chess game, or prices of goods if the interaction is a market negotiation), and another that assigns affective identities and behaviours to individuals involved. I will call the first subspace the “domain space”, $\mathbf{X}_x$, and the second the “social context space”, $\mathbf{X}_y$, and there must be a connection between the two. Linguistic labels for identities are normally negotiated at the very start of an interaction (e.g. “you’re a doctor and I’m a patient, and you are advising me to change my behaviour”). These labels have associated, culturally shared, uncertain, fundamental sentiments (E,P,A) in $\mathbf{Y}$, related to $\mathbf{X}$ through a somatic transform. They are related to the denotative state as well, but through the control mechanism, to which I return after presenting the formal semantics of *BayesAct*, in which conditional probabilities are encoded with each set of incoming links to a node. Thus, in the model shown in Figure 9, we require six such conditional probability distributions: $P(x'|x, a)$, $P(x, y)$, $P(a, f_b)$, $P(y|y, f_b)$, $P(\omega_x|x)$, $P(\omega_e|y)$. I will consider these in turn.

transition function: $P(x'|x, a)$. This can be learned from data, and has a parameter δ that governs the variance of the predictions. When δ is high, the predictions are weak (high entropy).²⁵ I will use $\hat{\delta} \equiv \delta^{-1}$ in this paper such that high values of $\hat{\delta}$ correspond to strong predictions (low entropy).

somatic transform: $P(x, y)$ and $P(a, f_b)$ are detailed in (Hoey et al., 2020), in which the somatic transform is modeled as a Boltzmann distribution. The somatic transform has a parameter γ that governs the variance of this transform. When γ is high, the variance is high and the transform does little for the posterior. I will use $\hat{\gamma} \equiv \gamma^{-1}$ in this paper. I will refer to $P(x, y)$ as including $P(a, f_b)$. This is consistent with an active inference view in which action strategies are computed using probabilistic *inference*, not *search* (see Section 4.3).

connotative dynamics: $P(y'|y, f_b)$ must specify a distribution (in sentiment space) over the action space of the agent. One method of doing this is by factoring Y into two pieces, one that is a measure of meaning “out of context” (based on memory or current context), and the other is a measure of meaning “in context” (taking the current actions of other agents and the self into account). In Affect Control Theory, the difference between this in- and out of- context sentiment is proposed as a measure of “unease” that an agent will attempt to avoid. Easing this difference is precisely optimizing the action space of the agent as the best action to take (connotatively) in a given situation. I will denote this best action to take as π , which can be defined as a distribution over its argument, such that $\pi(a)$ is the distribution over the denotative

²⁵Entropy is a measure used in statistical physics, and it describes the level of homogeneity of a distribution. In information theoretic terms, entropy measures the amount of information in a system that exists in a set of states x according to the probability distribution $P(x)$. Entropy is typically written as $S(P) = -\sum_x P(x) \log P(x)$. A low entropy system has a joint distribution over states $P(x)$ which is not dispersed evenly across all x , and so is easier to predict. Kahneman and Klein (Kahneman and Klein, 2009) refer to this as the “validity” of the environment.

action space, while $\pi(f)$ is the distribution over the connotative action space. *BayesAct* explicitly represents the uncertainties in these sentiments, and can be used to generate a joint posterior written as $P(y)$ as above. The connotative dynamics will be parameterized by α , which governs the precision of the connotative state predictions. This parameter is normally 1 in ACT (Heise, 2007), but only because there was nothing to compare it to. In this case, we see α as being defined relative to γ and δ . I will write $\hat{\alpha} \equiv \alpha^{-1}$.²⁶

Note that the observations are themselves noisy, and the agent will require a model of this noise, implying two further parameters at least (one for each $P(\omega_x|x)$ and $P(\omega_e|y)$). In what follows I will assume that these noise components may be unknown to the agent, but can be learned in the same way as $\hat{\alpha}, \hat{\delta}$ and $\hat{\gamma}$. In a more general treatment, these two additional parameters would move the analysis from three dimensions to five, which causes greater conceptual difficulties in gaining an understanding of the ideas. Further, it is possible to “collapse” these distributions into the state and meaning spaces, by integrating them with the $P(x'|x, a)$ and $P(y'|y, f)$. Once collapsed, $\hat{\delta}$ (for example) actually represents both $\hat{\delta}$ and $P(\omega_x|x)$ using $P(x'|\omega'_x, x, a) \propto P(\omega'_x|x)P(x'|x, a)$.

An agent in a group of three or more using such a model to guide its behaviour will be required to maintain a model of each other interactant in memory. In order to enable group behaviour, each agent must be able to make predictions about what other agents will do *to each other*, and a way of interpreting their interpersonal actions. I will assume that each *BayesAct* instance includes a connotative and a denotative model of one other agent, as well as a connotative model used for the group. Other, significant agents, may also require a denotative representation, but others will be relegated to the group as a whole. This allows agents to rapidly switch interaction patterns “on the fly” during a group interaction.²⁷

Emotions may be relayed through Ω_e , and give indications of differentials as indicated above. I assume that there is some mechanism for converting the emotional differential over Y to an emotional signal for another agent (his Ω_e). As in (Asghar and Hoey, 2015), I can assume that the differential is automatically conveyed, and so is impervious to manipulation by the agent, although we know this not to be the case.²⁸ Two remaining pieces are the discount factor and the utility function (not shown in Figure 9) which turns the Bayesian network into a decision network. The discount factor governs the search space and will be discussed in Section 4.2. The utility function is defined in *BayesAct* exclusively over the denotative state, and can be viewed as an integral part of that state (i.e. we can infer it from data). The definition of rewards (utility) for specific outcomes is a denotative labeling process that may also carry connotative meanings,

²⁶In fact, the α in (Heise, 2007) is the $\hat{\alpha}$ I am using in this paper, but this is not significant as it is not used in ACT.

²⁷I imagine implementing this using a set of “swap” bays (in software, also called “slots” or “schema”), and some storage systems that agents can use to keep information on other agents, such as their social identity and history to some degree. The schema are populated by instances from the storage system, and contain “maps” of other social individuals as distributions over Y space. *BayesAct* agents also have the ability to model an interaction *between* two such schema. In fact, the model of all interaction can be written as an interaction between schema that can be populated by data. If it is used to model an interaction between the *self* and some alter, then one of these schemas is drawn from a more consistent (longer-lasting) distribution over a similar sentiment space, but one that encodes the emotional history of the agent as sets of social identities that are more or less salient (MacKinnon and Heise, 2010). These schema may also include connections to larger such schema describing the institution or social group identity (Hoey and Schröder, 2015). Finally, such schema may be converted into an equivalent deep Bayesian model, as the one I have described. Structure in the schema is directly translatable to the Bayesian network (Koller and Friedman, 2009).

²⁸Relaxing this assumption requires a further complication of the search space (see Section 4.2).

but we do not model those here, see (Asghar and Hoey, 2015). Nevertheless, the optimization in the model attempts to maximize this utility function. It does so by predicting the future in X , but also predicting the future in Y , which will help it to establish cooperative solutions. The reward is defined over the denotative space exclusively, and may not arise “out of thin air” (a.k.a. be specified by some external control mechanism), but may be defined by the expected states to be visited (Friston et al., 2012) in a symmetric fashion where one reinforces the other. That is, expected states are there for a reason: because they are useful to an agent. The denotative association of those states with particular labels (e.g. the cost of an item) therefore creates a utility function implicitly (as preferences). I return to this reward function, described in terms of “values”, in Section 7.6.

4.2 Flashlight algorithms

One straightforward method for finding a *policy* of action in *BayesAct* is largely described in (Asghar and Hoey, 2015). The Monte-Carlo methods used in that paper search the state space of $\{X, Y\}$ looking forwards through time in an incremental fashion, and attempt to rank the possible actions a using the expected reward gathered. This forward search using $P(x'|x, a)$ is guided by the settings of this distribution by the connotative space, Y , however. Thus, the predicted future events will be less dispersed than they would be otherwise. That is, the connotative state *constrains* the denotative predictions to only lie in certain directions. Again, we return to the boys with the flashlights, and associate the width of the beams with the precision of the connotative predictions for each boy: more connotative precision means a narrower beam. A more dispersed somatic transform implies a larger target (the totality of the social beam), and a more dispersed transition function implies a less predictable target movement. Thus, the parameters $\hat{\alpha}$, $\hat{\gamma}$ and $\hat{\delta}$ are jointly defining a particular trajectory through the search space. Different settings of these parameters will imply different search strategies. When $\hat{\alpha}$ is large and the other two small, then our hero is seeking large targets moving unpredictably with a narrow flashlight beam. When $\hat{\gamma}$ is large and the other two small, he is seeking small targets moving unpredictably with a wide beam. Finally, when $\hat{\delta}$ is large and the others are small, he is working out the optimal trajectories of these rather predictable target movements.

When computing reward in such a partially explored search tree, the agent can set a *discount factor* which allows it to re-weight expected rewards based on their depth in the search tree. The discount factor is somewhat under the control of the agent. However, when exploring the search tree in *breadth*, an agent can use more discounting, as he does not even plan to explore the longer-term effects. When exploring in *depth*, an agent must use less discounting, otherwise what’s the point of exploring in depth? In Section 9, I return to the discount factor when describing one important such factor in economics which is the *inflation* (corresponding to a small discount factor, i.e. lots of discounting, making immediate rewards more important than long-term rewards and convincing people to buy now rather than wait).

All of these pleasant stories about boys and flashlights can be directly related to settings of the search algorithm in the very large space of possible futures. Consider the simplex shown in Figure 10. At location D , the search space is somewhat valid, and the flashlights are somewhat dispersed. Short and perhaps medium-term goals are “illuminated” by such beams, and a reasonably rational solution can be worked out.

Locations B and C have the same denotative precision, but in C it is provided mostly by $\hat{\alpha}$,

Figure 10: Simplex with four key locations discussed in the text.

and so will be highlighting (searching) parts of the state space corresponding to social rewards, such as cohesion. These rewards are ones that typically lie deeper in the search tree, and are typically much larger (e.g. the long-term rewards from a trusting and successful partnership). Thus, at C , the search focus will be more in the direction of these longer-term larger, deeper rewards, and these longer-term payoffs will be found. The discount factor here must be set high (little discounting) so that these future rewards look meaningful enough. These are large, deep targets (denotatively). The appearance of large, deep, targets is thus a by-product of the use of $\hat{\alpha}$ to afford precision. A diverse social world (as ensured by low $\hat{\gamma}$) will only lead to solutions if there is trust between agents. Such trust usually engenders larger, longer-term rewards.

At B , on the other hand, it is $\hat{\gamma}$ that is providing the connotative precision, and so determines the direction of the flashlight beam towards more immediate affective meanings, thus seeking rewards lying less deep in the tree. These are small, shallow targets (denotatively). Notice that the connotative precision here is the same as in C , so the flashlight beam is the same “width”. One (B) is “tuned” to small shallow rewards, while the other is tuned to large deep rewards. The appearance of small, shallow, targets to agents at B is a by-product of the use of $\hat{\gamma}$ to afford precision. However, a homogeneous social world at the level of meaning is required in order to make them appear.

At location E , the connotative space is largely rendered useless, and a wide flashlight beam results. Since it is so wide, it can only search very shallow depths, and will only find these shallow rewards. This is the standard reinforcement learning paradigm. At this point, some artificial exploration bonus must be introduced since there is no guidance from the connotative system. Normally, some vague references to an emotion (such as optimism) can ensure this exploration, but really this is moving the agent away from the point E by increasing $\hat{\alpha}$ or $\hat{\gamma}$. See further discussion in Section 8.

Finally, location A is the trickiest one as it has reasonably high $\hat{\gamma}$ and $\hat{\alpha}$ but very low $\hat{\delta}$. Thus, the environment is ambiguous and unpredictable, but social cohesion is strong, both in terms of short term meanings *and* in terms of longer-term social goals. Thus, the beams are narrower

than at D , but illuminate both short and long-term goals. Discount factors would need to be set according to the longer-term goal. Since $\hat{\delta}$ is low, however, expected rewards from long- and short- term goals are equalized to a certain degree, since the agent is more uncertain about it.

The agent may also perform a search through futures in the connotative space, guided by its predictive model therein and its coupling to the denotative space. In fact, both searches are carried out simultaneously and together, but could be done separately if needed. This type of search, coupled with the ability to manipulate emotional signals, may lead to devious (or helpful) trajectories through the emotional space. Humans are, of course, remarkably deft at doing this, but the additional complexity posed by this ability will only serve to cloud my analysis at this point. Nevertheless, such emotionally manipulative behaviour will cause variation in the social structure along the simplex, possibly moving some sub-groups off of it.

Time also becomes a factor in the search deliberations. The algorithm used in Asghar and Hoey (2015) is an any-time algorithm, so computes the best possible solution in the allocated time (basically until the next event happens, so requires an external clock). When the action distribution gets sufficiently un-dispersed, motivation to act will be stronger. However, the more time an agent is given to make a decision, the more of the search space he will explore, regardless of his search strategy. This allows agents to “break free” from the connotative restrictions and to discover new denotative dynamics. These discoveries are naturally integrated back into the denotative state space if possible, and can have a large influence on the future course of interactions in the social network. For example, diffusion of this innovation (discovery) will take place in some way that is dependent on the network structure (Lazer and Friedman, 2007) and also in some way on the connotative meanings distributed or created in the social group (Deffuant et al., 2005; Goldberg and Stein, 2018).

Now I can explain how the denotative social context space is related to the denotative domain space. This occurs through the action distribution and the ACT control principle, which, based on the identity context, generates a distribution in sentiment space for behaviour and identity, which “reveals” a set of denotative actions and identities that are affectively consistent with the context. This distribution over the space of denotative actions is then combined with the prior over denotative state, with the transition function, and finally with any observations made, to yield a final posterior over that same action and identity space. Only those actions which are both affectively consistent with the identity context, and also consistent with things that can be done in the real domain, will be revealed as possibilities for the agent to explore. This exploration consists of projecting those actions forward in time, and analyzing what outcomes will result. This analysis continues for as long as possible, until interrupted by a following event, and the most promising looking action can now be selected. Thus, the choice of action implicitly links the social context space with the domain space. Note that there may be other connections, for example through observations of identity, or expectations about domain space given identity. This type of domain knowledge may be specified in the POMDP as a prior.

For example, suppose my wife doesn’t get the job she applied for. In the domain space, this impacts finances, leisure activities, etc. However, in the social context, the only actions available to me are to *console* or *encourage* my wife “*No worries, we’ll be fine - you’ll get the next one, I know it*”. It is possible for me to *denigrate* my wife “*I knew you wouldn’t get it, you just aren’t good enough*”, and I may even know that, given her competitive personality, doing so will spur her to try even harder, therefore alleviating the domain problem (e.g. finances). However, given our positive relationship, this action would likely not even appear as a possibility to

me. Therefore, when combining my affective expectations for my own actions (*console*), with my denotative expectations (possibilities, really, *console* or *denigrate*), I end up with only one action to explore, *console*, and the problem is solved before I start. Perhaps there are a few ways for me to console her, though. I could buy her flowers, or make her dinner, or just sit and talk to her. All of these actions remain affectively consistent, and, if given enough time to do so, I may consider each in turn, evaluating what I think may be likely outcomes.²⁹

Random sampling (e.g. Monte-Carlo) methods suffer when scaled up to higher dimensional state spaces, as the number of samples required to maintain a reasonable posterior becomes intractable. This is certainly true for invalid environments, but may easily be true for even moderately sized valid environments. An alternative is to attempt to directly compute the posterior by approximating it with a class of functions that can be fit to the data. I turn to these methods next.

4.3 Free Energy and Active Inference

The complexity of an agent’s environment, defined as the total number of configurations accessible to the agent, is related to the *free energy* (MacKay, 2003). A configuration is one potential arrangement of the state of the world in the future. The accessible configurations are those that the agent can achieve as a function of its actions (and possibly exogenous events including the actions of other agents). The accessible configurations are defined by (and define) the agent’s *econiche* (Bruineberg and Rietveld, 2019). This is the set of dynamic configurations of other agents and objects in an agent’s environment. The stability of these configurations is crucial to the establishment of a long-lasting equilibrium or steady state.³⁰ Each agent must be aligned with his *econiche* (and thus to all other agents and objects in it) to achieve an equilibrium in which the number of expected configurations is minimized (or at least the probability over them is maximized, and the free energy is minimized). An agent wishing to assess the free energy and to use it as a guide for action needs to model the *distribution* of the free energy (so it can take a derivative and know which way to shift its model), which translates to the relative probabilities of all the future configurations of the world. Whether these configurations are known or not, a successful agent must predict how likely each one is.

As an agent’s world becomes more complex (e.g. with the addition of other social agents), this true free energy becomes intractable to model within an agent’s resource bound, and so the agent must approximate. The agent, primarily interested in evaluating how surprising outcomes will be, will need to average over all possible future configurations, something that it cannot reasonably do. It can, however, construct something that is always *bigger* (more surprising) than the surprise of the environment, known as the *variational free energy*. An agent’s *variational free energy* is an internal measure of how well its (approximate) model fits the real world, combined

²⁹In some marriages, none of these has even the remotest possibility of solving the problem because *consoling* is interpreted differently by the receiving party (perhaps as *condescending*). Such a case requires alternation by one of the parties, and can be achieved through therapy or nihilation (Berger and Luckmann, 1967). The first attempts to re-wire one or both parties, the second simply changes the goal for me (to one in which I don’t care about accepting blame, for example).

³⁰Some social environments may be extremely unstable, such as in warfare, but in such a situation, agents would be expected to behave in ways to re-establish an equilibrium. This may occur by attrition, for example, a famous example of which during the first world war also involved the development of localized cooperative *econiches* in which covert social signaling techniques were used to minimize casualties (Axelrod and Hamilton, 1981).

with the true degree of surprise (the true *free energy*). The idea of minimizing this variational free energy is the same as moving towards a state of true free energy which is minimal, and to a model that is the best possible abstraction of the real world given the resource bound, and is consistent with the econiche in which it is embedded (Bruineberg and Rietveld, 2019; Friston, 2010). The success of this model can be measured by how much the agent is surprised by its environment, measured by the *variational free energy*. A very well matched agent will not feel surprised very often, whereas a mis-aligned agent will often feel surprised. As surprise is a key factor in survival, the models have an intuitive evolutionary argument to support them. In fact, even evolution is thought to follow a similar optimisation of surprise, leading to species that are better and better adapted to their econiches (Badcock et al., 2019).

Policies of action can be computed using this variational free energy approach, and yield a quite general class of normative model of intelligence as based on *active inference* (Ramstead et al., 2019). This class includes solution techniques that can be used on the type of deep Bayesian model I am introducing here. The computations of a policy of action resolve to inferences about model parameters, coupled with inferences about expected futures. Since distributions over model parameters (over models) encode variational free energy, the resulting inferences give rise to policies that implement the same type of “flashlight” algorithm described in Section 4.2. The primary optimization is usually termed the *free energy principle*, which states that agents try to maximize control of their perceptions. That is, agents use their *output* to try to control their *input*, and only indirectly have effects on the world itself. This is fundamentally different than the traditional approach in artificial intelligence and engineering of mapping from *input* to *output*.

Moutoussis et al. (2014b) describe a free energy model of interpersonal inference (also described in Section 8) which describes in great clarity how an individual agent will minimize free energy of policies using a model of other’s behaviour based on *esteem*. The formulation of the active inference largely carries over to the deep Bayesian model I am considering, but Moutoussis et al. (2014b) make a number of simplifying assumptions that are removed in *BayesAct*. In particular, they assume that 1) esteem is a fixed additive function with no dispersion; 2) relationships between esteem and outcomes are fixed and defined based on simple heuristics (e.g. “good people do good things”); and 3) that self-beliefs are used exactly to model others. The dynamics function in *BayesAct* (characterized by dispersion α removes assumption 1, while impression formation (α again) gives explicit meaning to the relationships between beliefs about self and other, removing assumption 2, and the somatic transform (characterized by dispersions γ) removes assumption 3. That is, esteem will be added to expected reward through weightings based on precisions of self and other estimates, and of their expected connotative dynamics. I can therefore associate Moutoussis et al. (2014b)’s *social Pavlovian bias* with the connotative dynamics of *BayesAct*, and show explicitly how shared representations of self and other are propagated and maintained. The explicit modification of reward functions in Moutoussis et al. (2014b) is replaced by the flashlight algorithm, which places more weight on socially acceptable outcomes *a priori* (before any interaction). That is, attention in forward policy search is driven by connotative dynamics, leading to outcome distributions (and therefore rewards) focussed on the outcomes so predicted. Moutoussis et al. (2014b) also show a mapping to different brain regions, with the prefrontal cortex representing beliefs, the premotor cortex and striatum, actions, the VTA/SN, precision, and the inferotemporal cortex, observation. However, note that these relationships may hold at a different level in brain function (see Section 5.3).

I now derive the simplex in Figure 2 from free energy principles, for a set of agents in a group who all use the model in Figure 9.³¹ I will start with the free energy of the *whole group* of N agents considered as a single agent. I will denote the i^{th} agent's parameters as θ_i and the parameters of the whole group as $\theta \equiv \{\theta_1, \dots, \theta_N\}$. Thus, the task of the group at time t , given data as observations of emotional (e) and other (o) data $D = \{e, o\}_t = \{e_1, e_2, \dots, e_t, o_1, o_2, \dots, o_t\}$ is to compute

$$P(\theta|D, \mathcal{H}) = \frac{P(D|\theta, \mathcal{H})P(\theta|\mathcal{H})}{P(D|\mathcal{H})}, \quad (2)$$

where $\mathcal{H}$ is the hypothesis space (the modeling space as defined by the *BayesAct* model, for example). The graphical model for these agents is shown in Figure 11(a), shown with a single latent variable $\mathbf{Z}$ for the entire group. The difficult part here is the evaluation of $P(D|\mathcal{H})$ since it involves a summation over all values of θ . Further, each of these terms involves sums over the hidden variables $\mathbf{Z}$.

Figure 11: Simplified factor graph (Bayesian network) representing (a) a group with latent variables $\mathbf{Z}$; and (b) the group without agent i with state $\mathbf{Z}_g$, and one representing agent i with state $\mathbf{Z}_i$.

The free energy for this group model can be approximated with a variational function $Q(\theta)$, and the variational free energy, $\tilde{F}$, can be written as

$$\tilde{F} = \int Q(\theta) \log \frac{Q(\theta)}{P(D|\theta)}. \quad (3)$$

When Q and θ are chosen such that $\tilde{F}$ is minimized, then θ will be the best predictor function for this domain and agent combination (MacKay, 2003). The overall minimization process is normally handled by choosing Q for some fixed (current best-guess) θ , and then optimizing θ with respect to that “discovered” Q , and repeating this process until convergence to a local minimum. The key idea is that by choosing the Q function appropriately, a minimization over model parameters is possible based on it, and this minimization will not leave our parameters any worse off as far as their relationship (fit) with the data goes. In many cases, Q may be determined from the data, but in some it may only be possible over some parameterized subset

³¹If you, dear Reader, are less interested in the mathematical derivation, you may wish to skip ahead to the next section, where I proceed to analyze economic and social structures, which I can assure you is navigable without the derivations.

of the space of Q . For example, Q can be factored into parts corresponding to each parameter, and then each such factored Q can be minimized analytically one at a time, while keeping the others constant. The variational Q is in fact a “functional” which means a family of functions. The choice of this functional may be important, but typical solutions are found empirically or using very simple functionals. The method works by first “fitting” the functional Q to the probability distribution being optimized at some “guess” for θ . and then optimizing θ over the function so defined. This process is repeated until convergence and may be superior at finding the global optimum

In order to move beyond the group to each individual agent, I will split the group into two parts. One singleton set contains the i^{th} agent, with parameters θ_i and latent variables $\mathbf{Z}_i$, and the other set contains $N - 1$ agents $\{1, \dots, N\} \setminus i$, with parameters θ_g and latent variables $\mathbf{Z}_g$. I will consider this second set of agents as a single agent as I did above for the whole group. The role of this single “group” agent is, in fact, taken by a single individual from the group at any one time, but the statistics of interaction of the agent in question with the *whole* group is what matters, as I will show. At this point, I encourage you to think of this as a *dyadic* interaction, but it can also be viewed as an agent-group interaction, or even a group-group interaction. It is known that group identity (connotative) emerges as associated with specific group labels (denotative), and that these labels, if applied to groups, imply an *a priori* bias towards that connotative identity in individuals (Ridgeway, 2019, p.133,p.148).

In this restricted case, each agent is then trying to optimize its own posterior density (including over its own action space),

$$P(\theta_i|D, \mathcal{H}_i) = \frac{P(D|\theta_i, \mathcal{H}_i)P(\theta_i|\mathcal{H}_i)}{P(D|\mathcal{H}_i)},$$

and the network model looks as in Figure 11(b).

I will now assume that the variational distribution for the group, Q , for any agent i , can be factored into a piece for the agent, $Q_i(\theta_i)$ and a piece for the group, $Q_g(\theta_g)$, such that $Q(\theta) = Q_i(\theta_i)Q_g(\theta_g)$. As explained above, a variational solution to this model will normally require some kind of iterative updating scheme like the expectation-maximization algorithm, which operates by optimising one parameter at a time, while holding the others fixed. This kind of iterative solution is achieved by factorizing the group into individuals optimizing their own Q functions, based on everyone else’s Q functions, assuming they are fixed. For the entire group I am considering, I am assuming that each agent can separately and independently minimize its own variational free energy.

What this implies is that the quantity each agent in a group of agents will be optimizing is

$$\tilde{F} = \int Q_i(\theta_i)Q_g(\theta_g) \log \frac{Q_i(\theta_i)Q_g(\theta_g)}{P_i(D|\theta_i)P_g(D|\theta_g)}. \quad (4)$$

Consider D , the total data generated by the agent+group. I will break this into three non-overlapping sets, $D \equiv \{D_o, D_i, D_g\}$ where D_i is the data generated during the interaction by the agent, i , while D_g is the data generated by the group, g , and D_o is the data generated by both simultaneously. The goal of the optimization is to get D_i to be interpretable by the group, to get D_g to be interpretable by the agent, and to get D_o to be interpretable by both. Both agent and group will be updating their models, θ_i and θ_g , respectively, during the interaction.

I will therefore simplify by assuming that each agent generates “its” data, then observes D_i and D_g , and then generates the shared data D_o . Then I can factor

$$\begin{aligned} P_i(D|\theta_i) &= \int_{\theta'_i} P_i(D_o, D_i, D_g, \theta'_i|\theta_i) \\ &= \int_{\theta'_i} P_i(D_o|D_i, D_g, \theta'_i, \theta_i) P_i(\theta'_i|D_g, D_i, \theta_i) P_i(D_i|D_g, \theta_i) P(D_g|\theta_i) \\ &= \int_{\theta'_i} P_i(D_o|\theta'_i) P_i(\theta'_i|D_g, D_i, \theta_i) P_i(D_i|\theta_i) P_i(D_g|\theta_i) \end{aligned} \quad (5)$$

$$\begin{aligned} P_g(D|\theta_g) &= \int_{\theta'_g} P_g(D_o, D_i, D_g, \theta'_g|\theta_g) \\ &= \int_{\theta'_g} P_g(D_o|\theta'_g) P_g(\theta'_g|D_i, D_g, \theta_g) P_g(D_g|\theta_g) P_g(D_i|\theta_g) \end{aligned} \quad (6)$$

where I have assumed that D_o is generated from updated models in agent θ'_i and group θ'_g after seeing D_i and D_g . Further, I’ve assumed each agent computes its own $P(D|\theta)$ without considering the other’s data. That is, $P_i(D_i|D_g, \theta_i) = P_i(D_i|\theta_i)$ and $P_g(D_i|D_g, \theta_g) = P_g(D_i|\theta_g)$. Putting Equations 5 and 6 back into Equation 4, and rearranging terms, I obtain:

$$\begin{aligned} \tilde{F} &= \int Q_i(\theta_i) Q_g(\theta_g) \log \frac{Q_i(\theta_i)}{P_i(D_i|\theta_i) \int_{\theta'_i} P_i(D_o|\theta'_i) P_i(\theta'_i|D_g, D_i, \theta_i)} \\ &\quad + \int Q_i(\theta_i) Q_g(\theta_g) \log \frac{Q_g(\theta_g)}{P_i(D_g|\theta_i) P_g(D_i|\theta_g) P_g(D_g|\theta_g) \int_{\theta'_g} P_g(D_o|\theta'_g) P_g(\theta'_g|D_i, D_g, \theta_g)} \end{aligned} \quad (7)$$

Now I will evaluate this free energy at a fixed point where $\theta'_g = \theta_g$ and $\theta'_i = \theta_i$, and in the particular case where $D_o = \emptyset$, ergo, the group and individual are at equilibrium. This means neither agent nor group changes parameters based on the other’s data. However, at equilibrium, where I am computing this equilibrium, it allows us to compute the integrals in closed form, and thus is a fixed point solution. Thus, in Equation 7 I can set $Q_g = P_g(D_g|\theta_g)$ and set the integrals over θ'_i and θ'_g in the denominators to identity (since one term picks out $\theta' = \theta$ and the other is $P(D_o = \emptyset|\theta) = 1$ they pick out the equilibrium point, which is the starting point).

With these assumptions in hand, I can rewrite Equation 7 as:

$$\begin{aligned} \tilde{F} &= \int_{\theta_g} Q_g(\theta_g) \int_{\theta_i} Q_i(\theta_i) \log \frac{Q_i(\theta_i)}{P_i(D_i|\theta_i)} \\ &\quad - \int_{\theta_i} Q_i(\theta_i) \int_{\theta_g} Q_g(\theta_g) \log P_i(D_g|\theta_i) P_g(D_i|\theta_g) \end{aligned} \quad (8)$$

The first term is the usual free energy for the agent, averaged over models of the group. However, assuming the group is stationary, then the free energy of the agent then resolves to its own free energy, which can be computed. The second term is the joint probability that agent i will be able to generate data D_i that is interpretable by the group, and that the group will be able to generate data D_g that is interpretable by the agent. This is taken in expectation over both models of agent and group parameters, Q_i and Q_g .

Note the symmetry in Equation 8, in which the dispersion of θ_i can be large if the dispersion of θ_g is small, and vice-versa, but both can’t be large or small at the same time. To return to our flashlights, if the target is large (the dispersion of θ_g is large), then the beam must be small (the dispersion of θ_i must be small. In fact, this symmetry is quite curious because it states

that individuals operating in the first regime will be well suited to interact with individuals operating in the second. That is, although they are doing things differently, they in fact are complementary. In this case, the targets themselves are also the inverse beams coming from the group. There is a tradeoff between the width of these beams, with insufficient density if the two are large, difficulty finding the other if the two are small, leaving the agents to find tradeoffs in the middle. The exact location of this tradeoff is then something that must be negotiated, and may easily involve stereotypes of status. It also determines the sets D_o, D_i, D_g defined above, since e.g. if the dispersion of θ_g is small, most of the data will be generated by the group, so $D_o = D_g$, and the group will be considered “high status”, while the individual “low status”. If the dispersion of θ_g is large, the dispersion of θ_i is small, and so $D_o = D_i$ and the individual is considered “high status”, while the group “low status”.

Focussing on the second term in Equation 8 only, I can expand out the logarithm and get two terms

$$- \int_{\theta_i} Q_i(\theta_i) \int_{\theta_g} Q_g(\theta_g) \log P_i(D_g|\theta_i) - \int_{\theta_g} Q_g(\theta_g) \int_{\theta_i} Q_i(\theta_i) \log P_g(D_i|\theta_g),$$

which I can optimize separately. Holding the agent fixed at θ_i^* and optimizing θ_g in the first, and holding the group fixed at θ_g^* and optimising θ_i in the second. Then, this is

$$- \int_{\theta_g} Q_g(\theta_g) \log P_i(D_g|\theta_i^*) - \int_{\theta_i} Q_i(\theta_i) \log P_g(D_i|\theta_g^*),$$

Now, $D_g = \{d_{g0}, d_{g1}, \dots, d_{gN}\}$ and $D_i = \{d_{i0}, d_{i1}, \dots, d_{iN}\}$, which means, assuming all the data are independently and identically distributed given each model, we can write this as

$$- \sum_j \left[\int_{\theta_g} Q_g(\theta_g) \log P_i(D_{gj}|\theta_i^*) + \int_{\theta_i} Q_i(\theta_i) \log P_g(D_{ij}|\theta_g^*) \right],$$

I will assume that Q_g is a “hat” function which has constant probability over $[\mu_g^* - \sigma_g^*, \mu_g^* + \sigma_g^*]$ (so $\theta_g^* = \{\mu_g^*, \sigma_g^*\}$). Similarly for Q_i : replace all g subscripts with i . Next, I assume that P_i and P_g are normal distributions with parameters $\theta_i = \{\mu_i, \sigma_i\}$ and $\theta_g = \{\mu_g, \sigma_g\}$, and then the integrals can be done analytically to yield, for each data point, a contribution to the overall free energy of (note the extra negative sign that came from the $\log(Normal)$ distributions)

$$\frac{(\mu_g + \sigma_g - \mu_i^*)^3}{3\sigma_i^{*2}} - \frac{(\mu_g - \sigma_g - \mu_i^*)^3}{3\sigma_i^{*2}} + \frac{(\mu_i + \sigma_i - \mu_g^*)^3}{3\sigma_g^{*2}} - \frac{(\mu_i - \sigma_i - \mu_g^*)^3}{3\sigma_g^{*2}}. \quad (9)$$

To plot this, I assume we are at equilibrium, which I set at zero, so that $\mu_g^* = \mu_i^* = 0$ and $\sigma_i^* = \sigma_g^* = 1$ and I obtain two terms as follows

$$\left[\frac{(\mu_g + \sigma_g)^3}{3\sigma_i^{*2}} - \frac{(\mu_g - \sigma_g)^3}{3\sigma_i^{*2}} \right] + \left[\frac{(\mu_i + \sigma_i)^3}{3} - \frac{(\mu_i - \sigma_i)^3}{3} \right]. \quad (10)$$

The first is the negative log probability (free energy) that the group will align with the agent, which will be lower (more probable, lower free energy) if the agent is more “flexible” (can show a self the group will like, blue line in Figure 6). The second is the negative log probability the

agent will align with the group, which will be lower if the agent is more precisely defined (i.e. more “findable”, red line in Figure 6). Although in this case it is simply because I assumed we were at equilibrium, but in general will be true because the group can more easily “place” the newcomer.

Figure 12: Views of the minimum free energy points along each ray from the origin. (a) and (c): Axes are $-\log(\text{parameters})$, but correspond in scale to coherence ($-\log(\hat{\alpha})$), dependence ($-\log(\gamma)$), and validity ($-\log(\delta)$). I plotted the negative logarithm of each parameter, and colored the surface in (c,d), for visibility only. (b) and (d): Axes are the raw variance parameters (so larger is more dispersed, more freedom). The triangle in (d) is a sytlistic representation of a potential range of workable settings, lying in a simplex-like shape.

Now I remind you that the first term in Equation 10 in fact represents the entire group of $N - 1$ individuals. Therefore, by weighting the two terms equally (black line in Figure 6), I have made an implicit assumption that the group is fully connected, so that there are $N - 1$ terms like the first in Equation 10, and $N - 1$ terms like the second, that is, the individual interacts with everyone. This is not likely to be the case, however. What is more likely is that the newcomer interacts with only a dozen colleagues and managers, so his influence on the group is small. If we approximate this linearly and weight the first term in Equation 10 by N , then the dashed curves in Figure 6 result. The optimal free energy has shifted *outwards*, and more individual flexibility is called for in order to integrate the individual into the group.

Nevertheless, the individual is going to have to change more than the group, as the weight of

the population is in their favor (he is outnumbered). Ample evidence does suggest that social learning takes place (i.e. adopting the practices of the group) in proportion to the “*frequency of cultural traits around them*” (Heinrich, 2016, p.63fn4). That is, the more often they run into some group particularity, the more likely they are to adopt the practice, and integrate into their world model. However, if the individual’s parameters are substantially mis-aligned in general with the group’s but aligned with some sub-group’s parameters (setting of coherence, dependence and validity/ideology/moral matrix), then if the social network is constructed in such a way that this individual is mostly interacting with the sub-group, then these mis-aligned values may be strengthened within the sub-group. Should the group become large enough, or socially organized enough, their skepticism may be able “*to offer a challenge to the upholders of the ‘official’ tradition*” (Berger and Luckmann, 1967, p.121). This challenge may be handled by merger into the main institution (*internalization* by the sub-group of the primary group), which then enriches and differentiates this tradition, or by segregation of the skeptics, a process of *objectivation* that possibly includes dehumanization (change of their agreed upon assigned identity). Finally, the sub-group may gain sufficient strength to trigger a revolution, in which case the existing traditions are thrown away and replaced with the new ideology, and the sub-group *externalizes* to the group, society is produced by this sub-group, who define the new reality (Berger and Luckmann, 1967).

Equation 10 is the free energy for a one dimensional parameter space, under certain assumptions. In the three dimensional space shown in Figure 8, we can imagine this free energy curve, as shown in Figure 6, varies along any ray emanating from the origin, and that the minimum point defines the surface of the “simplex”, which is therefore revealed to be more of a “dome” shape (assuming radial symmetry). I therefore plot the simplex by seeking the minimum free energy along each ray from the origin. Plotting this as a function of $-\log \theta \propto \theta^{-1}$ yields Figure 12(a), with an interpolated, smoothed version in (c). Figure 6(b) and (d) are the same plotting θ directly. One can see it is *not*, in fact, a simplex. Nevertheless, recall that a social system becomes increasingly difficult to arrange as you move out along any dimension of freedom, and thus the actual range of operation of these parameters is likely to be relatively small, centered around a region in center of the minimum free energy manifold, shown stylistically as a triangle in Figure 6(d).³² Further, the simplex gives us a convenient way to discuss the manifold shown in Figure 12, so long as we remember that in practice it has this particular scaling. Note that the minimum free energy goes to ∞ as any parameter drops to zero, as Equation 10 blows up. It is therefore more difficult to plot the simplex as a function of the inverses of the parameters (Figure 6(a) and (c)).

There are numerous assumptions and shortcuts in the above analysis, but my objective was to derive a first approximation to this function. The assumption that group and agent are fixed are quite restrictive, and the analysis above simplifies the simultaneous change of agent to group and from group to agent (which change may be modulated by the “gravity” of the group) by using the symmetry of the problem. This simplification allows me to hold one agent fixed and modulate the other (or hold the group fixed and modulate the agent). Nevertheless, any more complex and reciprocal change would be characterized by the same equations, except with perhaps a coordinate change. Thus, I have proceeded with loss of generality only in the assumptions made (such as $D_o = \emptyset$), but relaxing these assumptions would generate multiple interesting avenues for future work.

Generality is also reduced by the fact that I left out external forces altogether. However, such

³²The triangle is inverted from Figure 2 because this plot is in the inverted parameters.

forces could be added to the equations above as an additional element that could be modeled, and would share responsibility for D (along with the agent and group). Adding such elements may skew the overall structures shown in Figure 12, but will not change the core ideas I am presenting. This does, however, remain for a topic of future research.

5 Social Structures

Now I examine the types of social structures that are expected to arise at different points across the simplex discussed in Section 2.3, and thus to the parameters, $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$, of the Bayesian model.³³ I will do so in two ways. First, by showing how the parameters in particular settings, must necessarily lead to certain social structures. I show a close correspondence between the pattern of social structures on the simplex and classic results in sociometry. I will return to this relationship in Section 7. Second, I go the other way and show how particular social structures (those at the apices of the structural space of triads) would remain stable under only certain parameter settings in the deep Bayesian model. By demonstrating that these two relations hold, I infer that the parameters of the deep Bayesian model will induce certain predictable social structures, and that certain social structures will be more easily at equilibrium for certain parameter settings of the Bayesian models. That is, I will show that the social structures and the deep Bayesian models are isomorphic in some sense.

This then gives insights in two ways. First, it allows a better understanding of the parameters of the model and their meaning in the real world. This will, in turn, allow the design of more effective agents that are socially and emotionally aware. Second, it gives a new angle to look at the diversity of human cultures, by relating structures observed to parameters governing uncertainty in a deep Bayesian model. This new angle can be used to predict the impact of changes to the social structure which are imposed externally to an agent (e.g. a natural disaster or a revolution).

In this analysis, I will focus exclusively on triads.³⁴ Further, I will assume very limited agents in triads who only have the capacity ($|\mathcal{X}|$) to handle one other agent. This may seem restrictive to you, dear Reader, but I will attend to your concerns in Section 5.3. I will use a maximum state space size of $|\mathcal{X}| = G \times N/2$, where N is the size of an agent’s network, and G is the required state space size to model a single other agent.³⁵ More trickery!: this is an obfuscated way of saying that each agent can only model a maximum of one other agent in a triad, but nevertheless may still be useful for larger networks. Lastly, I will ignore the reward in this analysis, but will assume that it is there at some points (even more trickery!).

³³As we have seen, the parameters are in fact the *inverses* of the variances in the model, and so large values of $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ correspond to undispersed distributions in their respective dimensions. That is, $\hat{\alpha} = \alpha^{-1}$, $\hat{\gamma} = \gamma^{-1}$, and $\hat{\delta} = \delta^{-1}$.

³⁴It is thought that the step from dyads to triads is one of the crucial ones for the establishment of social structures (Ridgeway, 2019), as it requires agents to think about other agents’ beliefs about them (second order beliefs). I have argued that these second order beliefs cannot be held denotatively due to resource constraints, and therefore that they must be held connotatively. The result is the establishment of beliefs about the *bounded competence* (Ridgeway, 2019, p.44) of others, referring explicitly to the “width” of the flashlight beam encoded by the group: the competence of others is evaluated *relative* to the illuminated region of the denotative search space, and so the competence of others can only be bounded.

³⁵This assumes all the other agents are equal. If agent i requires model size G_i , then $|\mathcal{X}| = \frac{1}{2} \sum_{i=1}^N G_i$. The factor of $\frac{1}{2}$ models the fraction of other agents that can fit in within an agent’s resource bound.

5.1 From structures to settings...

I have explained how the *BayesAct* model can be described with three parameters on a simplex, $\hat{\delta}, \hat{\alpha}$ and $\hat{\gamma}$. At any point on this simplex, we retrieve a value of these parameters, and *Bingo!* we have a *BayesAct* model. Such a model will match well with models like itself, and potentially others. Let us look therefore more carefully at these parameters, and find out what kind of model each will yield. From there, I will show how these kinds of models lead to particular structures.

Consider $\hat{\delta}$, the degree to which an agent relies on its denotative world model. As discussed above, an agent with high $\hat{\delta}$ will do well by himself, not having the resources to admit others. This agent will construct a decision-theoretically rational model of its environment, and will therefore be *rational* in this sense. I can therefore attribute $\hat{\delta}$ to the label “rationality”, and postulate that the amount of rationality (to be defined more precisely) in a given structure is directly proportional to this parameter. The amount of rationality can be directly inferred from the difference between the amount achieved and the expected value. This will work in very small groups if everyone has the same denotative model. In larger groups, the coordination required is too difficult, and an authority must be defined to enable this coordination.

Next, $\hat{\alpha}$ can be equated with “transitivity.” As $\hat{\alpha}$ is a model of connotative dynamics, a shared such model will ensure that the connotation of predicted choices of actions of others in a group will be consistent with the connotative predictions made by each agent. For example, each agent in a group may expect the others to cooperate. Thus, agent C will realize that his cooperation with agent A will be reciprocated by agent B. With different types of agents (i.e. different social identities or status characteristics), this induces transitive relations in human networks. Division of labour becomes possible with increasing $\hat{\alpha}$ as each agent can rely more fully on the others. Nevertheless, agents operating in this highly transitive world will be capable of adapting to new environments fairly easily, and will be tolerant of diversity and invalidity (ambiguity). This is because they rely on their connotative predictive model, and discount the emotional and concrete evidence. Further, they will not require authority as these transitive structures “self-police.” Disagreements are resolved restitutively, which does require enforcement, but it is imaginable that this enforcement be privately agreed upon by members. I therefore postulate that the parameter $\hat{\alpha}$ in the Bayesian model is a “transitivity setting”, and I will take the number of transitive relations to be directly proportional to this parameter.

Finally, consider $\hat{\gamma}$, which I can equate with “republican equality”. A group with high $\hat{\gamma}$ will have consistent interpretations of immediate events, but will not form a consistent dynamical model of others. This allows for agents to operate largely independently, with a small number of (denotatively) modeled identities. It calls for agents to be homogeneous, as they must share these interpretations. Thus it also calls for authority to define the ideal agent that all others must be homogeneous with, and to force agents to “fit in” this mold. The reason this equates to republican equality, is that to ensure homogeneity, some people have to be excluded, and thus all are susceptible to the arbitrary decision of the authority. In this case, the consistency between evidence and sentiment is thus guaranteed (vacuously), but agents must have the same interpretations of events. As their denotative models are the same, and they feel the same way about outcomes, their reward structures will naturally form to be the same (e.g., as the set of states they tend to visit), and therefore their homogeneity is defined by their consistent preferences, which leads to equality. A group with low values of $\hat{\gamma}$ will tolerate diversity, but this

leads to republican freedom (inequality). With different interpretations of events comes one-way relationships $A \rightarrow B$ in which A may very well feel that she is preferred by B (in that B would perform some act of preference towards A), although this is not the case. I therefore postulate that the parameter $\hat{\gamma}$ is directly proportional to the amount of republican equality in the triad.

Recall that the three parameters are not independent, but rather only admit workable solutions that lie on a simplex, which can therefore be described in two dimensions (Martin, 2009). This simplex structure appears triangular when viewed in two dimensions. However, it is the relationship of the three dimensions of variation to this simplex that is of interest, and therefore I maintain all three parameters and work in a three dimensional space. Nevertheless, these findings can be mapped into a two dimensional space when necessary for comparison to prior work.

With these associations in hand, we can compute the parameter settings that are best suited to modeling specific social structures formed by networks of individuals. I will do this by looking exclusively at triadic relationships, of which there are 16 types (Holland and Leinhardt, 1970; Feld and Elmore, 1982),³⁶ shown in Figure 13. These sixteen triads are connected by directed edges defining cooperative (friendly, affectionate) relations between individuals, considering all three individuals to be indistinguishable. They are numbered in Figure 13 using a three-letter code developed by Holland and Leinhardt (1970). The field of sociometry, now largely subsumed in the nascent field of “computational social science” (Cioffi-Revilla, 2014), was pioneered by (among others) Moreno (Moreno, 1934), and was based on questionnaires asking individuals in a group who they would “choose” (who they “liked”). These relationships were found to be asymmetric, and inductive of inequality, and the patterns so discovered could be linked to individual behaviours (such as running away from a group or switching groups). More recently, Burt (1984) argued for the use of the “important matters” question in which links in networks were established according to who one would discuss “important matters” with.

The three letter codes in Figure 13 encode the number of mutual $\bullet \longleftrightarrow \bullet$, asymmetric $\bullet \longrightarrow \bullet$, and null pairs $\bullet \quad \bullet$ (in that order), with additional letters (U,C,D,T) for situations that have some additional symmetry but the same census of pairs. While early sociometric studies were primarily interested in transitivity as a key driver of structure, Feld and Elmore, following the indications of Moreno (Moreno, 1934, 1947), recognized that inequality was also a significant and unaccounted for element (Feld and Elmore, 1982). Feld and Elmore plotted the sixteen triads in a triangular structure, with corners corresponding to the number of single, double and non- choices. This resulted in the structure shown in Figure 14 (a). I have plotted what is a triangle in (Feld and Elmore, 1982) in a three dimensional space, then selected a view point that is looking directly along the $z = 0$ plane (which is $\hat{\alpha} = 0$ in Feld and Elmore). The dimensions correspond to the number of choices made. Notice how the triads roughly form a triangle.³⁷

The triangular structure of (Feld and Elmore, 1982) fails to take into account that single-choice individuals may also be participating in a transitive relation, and so does not represent well the tradeoffs between inequality and transitivity. While they correct for this in (Feld and Elmore, 1982, fig 4, p83), shown rotated in Figure 14(b), we find a more parsimonious ternary structure by considering the transitivity dimension to not directly include intransitivities (as is done implicitly in (Feld and Elmore, 1982, Fig 3., p82; Fig. 4., p.83). Making this association, and

³⁶Modern sociometric analysis uses much more complex functions than simple “liking”, but here I find I can make my point using this most simple of graphs.

³⁷According to the scheme for measuring $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$ in a network I define later in the section, the data is shown in Table 1 and plotted in three ways in Figure 14.

Figure 13: Sixteen possible triads of relationships. The unconnected triads 003 (top left) also shows an alphabetic key labeling (A,B,C) the agents that I will use in the text.

Figure 14: Structures of triadic networks. The *BayesAct* interpretation of the three axes is shown in (c), while the other two show Feld and Elmore's (Feld and Elmore, 1982), with (a) from the original view (up the $\hat{\alpha}$ axis) and (b,c) from the view I have been using (John Levi Martin's (JLM) view), the view is from the $\hat{\delta} = 0$ plane at $\hat{\alpha} = \hat{\gamma}$ looking towards the origin, so the simplex looks like a triangle.

	triad	trans.	vac.	intrans.	F&E	ineq.	3-ineq.	payoff	4-payoff
		$\hat{\alpha}$	trans.		fig 4	(equal.)	$\hat{\gamma}$		$\hat{\delta}$
003		0	6	0	3	3	0	2/3	10/3
102		0	6	0	3	1	2	8/3	4/3
012		0	6	0	3	2	1	7/3	5/3
201		0	4	2	1	1	2	10/3	2/3
120U		2	4	0	5	1	2	10/3	2/3
120C		1	3	2	2	1	2	10/3	2/3
120D		2	4	0	5	3	0	10/3	2/3
021U		0	6	0	3	3	0	8/3	4/3
021D		0	6	0	3	1	2	8/3	4/3
021C		0	5	1	2	1	2	8/3	4/3
210		3	2	1	5	2	1	11/3	1/3
111U		0	5	1	2	0	3	9/3	3/3
111D		0	5	1	2	2	1	9/3	3/3
030T		1	5	0	4	2	1	9/3	3/3
030C		0	3	3	0	0	3	9/3	3/3
300		6	0	0	6	3	0	12/3	0/3

Table 1: Data used to create Figure 14. The “F&E Fig 4” shows the actual data in (Feld and Elmore, 1982, Fig 4.), whereas the columns labeled $\hat{\alpha}, \hat{\gamma}$ and $\hat{\delta}$ are plotted along those dimensions in Figure 14 in this paper. trans.=transitive, vac.=vacuously, intrans.=intransitive, ineq.=inequality

rotating the axes ($x \rightarrow z$, $z \rightarrow y$, and $y \rightarrow x$), I obtain Figure 14(c). The “new” triangle now has three bisector axes³⁸ as follows:

Figure 2	high	low	axis
G	$\hat{\gamma}$	$\hat{\delta} - \hat{\alpha}$	equality
A	$\hat{\alpha}$	$\hat{\gamma} - \hat{\delta}$	transitivity
D	$\hat{\delta}$	$\hat{\alpha} - \hat{\gamma}$	non-defined in Feld& Elmore (rationality)

I return now to transitivity, inequality and rationality and define methods for computing them from network structures.

- **transitivity** is measured in a triad in the classical sociometric literature using $6 - m$ where m is the number of definitely *intransitive* pairs of relations. A transitive relation is one in which, A chooses B, and B chooses C, and A also chooses C. An intransitive relation is defined as one in which A chooses B, and B chooses C, but A does not choose C. There are six possible such intransitivities in a triad. Logically however, transitivity is also a property of a triad in which *either* A does not choose B, *or* B does not choose C, and A does not choose C. In such a case, the premiss (that A chooses B and B chooses C) does not hold, so the conclusion must hold (vacuously). The vacuous transitivity, however, are not really transitivity, they are rather a lack of intransitivity, as discussed above. Thus, I define transitivity as simply the number of transitive relations (non-vacuous) that exist in the triad. This number will be handled by a similarly scaled value of $\hat{\alpha}$ in the Bayesian model.
- **inequality** is measured by counting the number of individuals who are chosen either twice (by both others) or by nobody in the triad. If everyone is chosen the same number of times, I denote this as a triadic relationship of equality (a prime example being the cyclic triad 030C, and the triad 111U, in both of which everyone is chosen once). Thus, the amount of *equality* in the triad can be measured with $3 - I$ where I is the number of double and non-choices. As I have equated the parameter $\hat{\gamma}$ with equality, I will use $3 - I$ as a measure of $\hat{\gamma}$. The astute reader of this and of Feld and Elmore (1982) will notice that triad 300 will be labeled highly unequal in this scheme. As I will discuss in greater detail in Section 3, this seems intuitively incorrect at first, until one realizes that such triads are *more* heavily constrained than those with more single choice nodes. The introduction of *identity* or of *rationality* as $\hat{\gamma}$ decreases (as inequality increases) forces things like a division of labour to make up for this increased inequality. In fact, this implies increased equality in the sense that each is equally tied down to their respective identities (high $\hat{\alpha}$) or plans (high $\hat{\delta}$). Although republican freedom is gained (republican equality decreased), negative or positive freedoms are decreased.

Thus, for triad 300, the solidarity (equality) is provided by $\hat{\alpha}$. People are tied down to their identities, have lack of negative freedom, but they gain freedom from other’s will and freedom of opportunity (even though this is, in fact, restricted by the identity maintenance imperative, although they cannot perceive that). For triad 003, solidarity

³⁸These are axes that lie on the simplex, and run from an apex to the center of the edge between the two other apices.

(equality) is provided by $\hat{\delta}$, and people are tied to a rational plan, have lack of positive freedom, but they gain freedom from unaccountable will and freedom to choose their own actions (even though this is, in fact, restricted by the rational plan, although they cannot perceive that).

- **rationality** requires a more thorough investigation. Recall that the $\hat{\delta}$ axis is defined as the *validity* of the environment. Valid environments (from the perspective of some agent, B) are ones in which the agent is able to predict what will happen next with a reliable estimate of risk, either because of environmental simplicity or agent expertise. How can network structure and validity be connected? One method is to consider what type of environment would be most conducive to the maintenance of which binary network structure. For example, the structure $B \rightarrow A$ is one in which B defers to A, but A does not defer to B. Consider a social dilemma in which the deferred-to agent (A) gets more resource than the deferring agent (B). Why would B want to stay in such a relationship? One reason is that B cannot estimate how likely it is that A will defer. As agent B’s connotative system then takes over, the positive action of deferring will be selected before the selfish option (assuming positive identity, such as “research participant” or “student” for B). From B’s perspective then, a situation of high invalidity could result in the stability of this structure. Even after one interaction, the validity is expected to increase, but here I consider only the single interaction, and statistics are gathered over many different instances of this couple, not over many iterations of a single pair of individuals. Thus, $B \rightarrow A$ could result from an asymmetry in validity between B (invalid) and A (valid), with the resulting pair operating in a moderate amount of validity. Similarly, a dyad with no link, $B \nleftrightarrow A$, has two selfish agents, a structure that could only sustain itself in a very valid environment. Finally, a dyad with two deferring agents could be the result of a very invalid environment for both.

For the purposes of this paper, I will characterise this dimension using a standard prisoner’s dilemma (Axelrod and Hamilton, 1981), in which two cooperative agents get 2 “points” each, while two non-cooperative agents get only 1 point each. Asymmetric cooperation yields 0 for the cooperator and 3 for the non-cooperator.³⁹ It is well known that rationality in this game implies non-cooperation, with all agents receiving an average payout of 2/3 across all games (this is what is achieved by the triad 003, as each agent only participates in 2/3 of the total games played). Perfectly cooperative triads (300) get an average payout of 4/3 as they each receive 2 with each play, and agents in the cycle triad 030C gets an average payout of 1. I postulated that, for triads, the parameter $\hat{\delta}$ is directly proportional to the inverse of this payoff (higher payoffs imply less rationality for the group), I scale it as $\hat{\delta} \propto \frac{4-R}{3}$ where R is the average payoff per agent in the triad playing a 0/1/2/3 prisoner’s dilemma game.⁴⁰

³⁹This two-agent game can be framed as a “giving” game in which each of two agents have a choice of either giving 2 points from a common pool (not owned by either agent) to the other agent, or taking 1 point for himself. The agents make their choices without knowing what the other has chosen.

⁴⁰My definition of rationality as purely decision theoretic leads to certain interesting predictions. For example, one would expect highly *valid* environments (due to either agent expertise or environmental predictability, see Section 2.1), to have very high $\hat{\delta}$ regardless of the uncertainty management techniques employed by agents that are involved. Therefore, a group operating in a valid environment can simply adopt a rational solution (which they can compute), and end up implementing triad 003 for example. However, a rational solution still may exist for a highly invalid environment, leading to the definition of this axis as decision theoretic rationality, not validity. The underlying meaning of the $\hat{\delta}$ dimension is that of *validity*, but for triadic structures this can be

Perhaps a simpler way to define irrationality is as the *cost* of performing a certain action, where I measure cost as the risk an individual is taking by doing this action, *relative to* the risk he would face if he did another action that is more according to the group’s prescription (social learning). That is, how much is an individual “sticking his neck out” by going with the group. In the prisoner’s dilemma, the cost, in this sense, of cooperating is 3, and so the irrationality is high, and $\hat{\delta}$ will be low. Similarly, defection has no cooperation cost, so irrationality is low, and thus $\hat{\delta}$ is high.

The result is shown in Figure 14 (c). Again, we see a very similar type of “planar” structure loosely arranged within a triangular boundary. A reasonable correspondence between this and the triangular structure of Feld and Elmore (corrected and rotated as in Figure 14(b)) shows that that “rationality” (my D axis) may indeed be a definition for the final bisector axes in the Feld and Elmore ternary structure (the other two being “transitivity” (my A axis) and “equality” (my G axis). The only difference between Figures 14(b) and (c) is that Feld and Elmore (b) consider a measure of transitivity that *subtracts* the number of intransitive relations. In *BayesAct* this is not necessary as intransitivity is described as a lack of all (including vacuous) transitivity. However, this means that in *BayesAct* I am counting the (negative of) the number of intransitive and vacuously transitive pairs (since the total must sum to six), whereas Feld and Elmore subtract an additional factor for the intransitive relations, but it’s not clear why.

I made some connections above in order to “align” the structural space of triads (with its properties of transitivity, equality, and rationality), to the space of uncertainties (the simplex). I assumed a simple alignment of transitivity= $\hat{\alpha}$, equality= $\hat{\gamma}$ and rationality= $\hat{\delta}$. However, one can use apply a number of other linear mappings, e.g., to align with axes of solidarity I review in Section 3. The main point is to demonstrate that roughly the same patterns can be seen from structural or from uncertainty viewpoints, each of which will be consistent with the other.

5.2 ...and back again (from settings to structures)

Here we consider the three apices of the simplex, with $\hat{\alpha} = \hat{\delta} = 0$ (dependence), $\hat{\alpha} = \hat{\gamma} = 0$ (validity) and $\hat{\gamma} = \hat{\delta} = 0$ (coherence), and argue how different structures naturally arise in this configuration of uncertainty management.

validity:

In this situation, the agent has a highly precise denotative model, but no social or group models.

Figure 15: Triad 003

Thus, the optimal triad will simply be as in Figure 15, each agent is isolated and does not

mapped to the payoff in a prisoner’s dilemma by considering the dyads involved to be acting in this way because of validity/invalidity. The end result is that, in this case, more *valid* dilemmas will yield *lower* rewards as agents will naturally try to optimize this non-cooperatively. The cooperative solution (low $\hat{\delta}$) is irrational in this case.

require any connotative input. Such an agent will only receive emotional signals interoceptively, denoting states of hunger and tiredness, for example. These weak signals may grow strong and change the denotative model, but largely the denotative solution will take over. The agent gets no emotional signals from other agents and so Y is “free floating” and will largely only get in the way (consume resources). Therefore, the best an agent can do is reduce $\hat{\alpha}$ or $\hat{\gamma}$, and would in fact prefer to reduce both since maintaining models at low values of these parameters implies high entropy model distributions, which contain little information and are easy to store. Therefore, we expect such a structure to occur at high $\hat{\delta}$, but low $\hat{\alpha}$ and $\hat{\gamma}$.

coherence:

Highly coherent agents have precise connotative distributions (prior to seeing any evidence).

Figure 16: Triad 300

These connotative distributions are in fact those that guide the agent towards cooperative behaviours, and so are beneficial to the agent with him actually knowing it. The key is that all agents in the group must share these same connotative distributions. As I have discussed, this is largely taken care of by socialization, identity acquisition, and learning in a stable group.

Therefore, for such agents, any “interference” from the denotative state, X , will be likely to disrupt the social balance held by this perfect triad. Therefore, it must be the case that either $\hat{\delta}$ is small or $\hat{\gamma}$ is small. But, if $\hat{\gamma}$ is large, then $\pi(a)$ will have to match $\pi(f_b)$ exactly (the connotative and denotative action distributions must agree), but this will be hard to do if the distribution over X is too dispersed, requiring each agent to reduce its denotative ambiguity and find better matching (more complex) models, and so $\hat{\delta}$ must increase as well. Similarly, if $\hat{\delta}$ is large, then the action predictions require a large $\hat{\gamma}$, otherwise they will severely corrupt the action distributions. This shows that $\hat{\delta}$ and $\hat{\gamma}$ both must be small for this structure, and 300 is the optimal one, shown in Figure 16.

dependence:

The final case has a strong somatic transform, and thus has connotative and denotative agree-

Figure 17: Triad 030C and Triad 111U with labeled nodes (see text).

ment, but has little connection to the real world or to others. Thus, such puzzling trios such as the cycle (030C) and the case of two “friends” of whom one is interested in a third, but the

third and one of the friends don't accept each other (Figure 17. Consider the cycle. It clearly supports irrationality (e.g. a money-pump), and can be achieved with little coherence as each agent does only two things (give and receive). Compared to 300, it requires less identity maintenance, as each agent does not need to be someone different in each relationship. The second case is similar, in that transitivity is low due to the missing link

I have therefore uncovered that there is a usable mapping from the three parameters governing a two-layer Bayesian model of uncertainty, and the triadic structures that arise given those types of agents. This will allow me to show how the dimensions of social structures found in (Martin, 2009) can be mathematically modeled with this two-layer Bayesian apparatus. First, however, I turn to the question of how these concepts will scale.

5.3 Smarter Agents

What if agent's capacity is increased, and so the environment validity increases. At a certain point, the agents will be able to better model each other, but will also form groups with other people, creating a more complex social network. When this occurs, then multiple triangles start to interact. However, as we know that each agent will have some parameter settings to manage his uncertainty, those parameter settings will "fit" better with others of similar settings. Further, the structures that form will be the ones most conducive to exactly that parameter setting. In each triad, I expect to see more of the types I have described above. Therefore, the resulting social network, for given parameter setting, will be the one that yields the larger number of such triads. These networks will be different as described in (Martin, 2009), and we expect to see web-like structures at high values of $\hat{\gamma}$, multi-hub like structures at high levels of $\hat{\alpha}$, and pyramidal structures at high levels of $\hat{\delta}$.

I can, to some degree, associate the parameter settings as a crude approximation to "culture", defined as the "social conventions that associate practices with meanings" (Goldberg and Stein, 2018, p.4). A critical part of this is identity, but the set of parameters as a whole defines identity "*types [which are] social products tout court, relatively stable elements of objective social reality*" Berger and Luckmann (1967, p.174). As discussed in Section 4.3, a new agent with a different set of parameters, if introduced to this well functioning social network, will be faced with a large degree of surprise, and will need to adjust parameter settings to match the group, or will have to change the group's parameter settings through any of a wide variety of means which go far beyond the scope of this paper. The result is, however, the key element of a shared model of connotation and of denotation. As I have explained, this shared model can be represented as a point in a three dimensional parameter space of a deep Bayesian model. It is of course possible that this is not a point at all for each agent, but rather a *distribution* over this parameter space, itself parameterized with yet another, higher level in the deep Bayesian model. It is a reasonable hypothesis given the current evidence that there are on the order of five such levels in the human brain (Smith et al., 2018). Regardless, we also have evidence of a three dimensional "emotional" space that is an agreed upon cultural qualia (Osgood et al., 1957). This emotional space can be mapped to one or more of these higher levels. In a formal Bayesian sense, and from the modeling perspective, any subset of these five levels may be collapsed by integrating out the un-needed variable (level). This integration yields another, slightly shallower, deep Bayesian model that contains elements at a certain level of abstraction that may not be exactly what is present in the human brain. That is, the elements that reach

the broadcast network or global latent workspace (VanRiullen and Kanai, 2021) and are brought into consciousness (and so are reportable on semantic differential scales) (Barrett and Satpute, 2013), may themselves be abstractions of what is present in the deep Bayesian model. It is one of these abstractions that we associate with the three dimensional structure of meaning exposed by Osgood (Osgood et al., 1957). However, my guess is that the human model has five or more “layers” (again, these will be loosely defined on the physical substrate of the brain, but more precisely defined in terms of the functionality of various networks in the brain), but that the same three-dimensional uncertainty matching problem is going on in an embedded and hierarchical fashion. That is, there is one pair of layers very low down in the brain’s processing hierarchy (very close to sensation and motor action) in which “denotative” state corresponds to actual physical signals, while “connotative” state is an endocrinological response to threat or reward, for example. Deviations in the denotative state from expectations create “deflection” at this level of connotation, but the effect is the same, the connotative level “takes over” when denotative predictions fail (see the escalator example in Section 2.1).

Connections to the “population ecology” approach are by associating the parameters space in *BayesAct* to be what Hannan and Freeman (1977) call a “blueprint” for a set of rules or procedures. In this view, such a blueprint is defined by three factors, which mirror the three dimensions in Figure 2. First, the formal structure (“written rules of operation”: $\hat{\delta}$). Second, the patterns of activity (“what actually gets done by whom” : $\hat{\gamma}$), and third, the normative order (defining what is “right and proper” : $\hat{\alpha}$). The other concepts that have followed in this space could be used to structure investigations into how *BayesAct* agents form organizations that compete with other organizations on a market, and from there to the economics literature.

By defining a Bayesian model in this way (with only two layers), my aim was twofold. First, I wished to investigate the properties of the hierarchical model in general while connecting it to existing mathematical sociological models, which I did in this section. Second, I wished to expose the common ground between political and sociological models and this modeling of uncertainty, to which I turn next.

6 Psychological Economics and Political Sociology

This section could easily lie in the early part of Section 7, as Rousseau influenced many post-revolution thinkers, including Durkheim. In *The Social Contract* (Rousseau, 1762), Rousseau was the first (in modern times) to propose a social contract as a head of state. A contract that nevertheless must be defended by the people upholding it. A people with any kind of diversity then need to balance the three freedoms in the social contract so that the contract fits the people. Therein lies the difficulty, and Rousseau examines essentially the same 3D space as I have been describing (Figure 2), except framed in political economic terms. Rousseau’s focus is on a model of social order that includes three components, and he describes tradeoffs between the three components that are reflections of the same simplex. These three components correspond roughly to executive government (enforces laws), legislative government (enacts laws), and the judicial component (evaluates laws). Rousseau also describes three forms of government that can arise from placing emphasis on one of the three components: Aristocratic, in which a small group of “superior” beings rules everyone else (this category it turns out also includes Marxists); Monarchic, in which the people give all their republican freedom to a king; and Democratic, in

which the government rules and people have rights. These three components may be combined in a “mixed form” of government, such as that of the United States. The judicial is formed of the intellectual elite (judges), a supreme class of being that can make “correct” moral judgments. The legislative enacts laws to represent the people, and can be large or small. The executive is the president and cabinet, whose unaccountable will is held in check by the other two (at least that is the intent).

At any point in history, the current state (the current group of participants) can establish a *sovereign*, which I have been calling *the state*, using these three components, balanced at some point on the simplex. Three institutions are required: legislative (the law), republican (the people) and constitutional (the president). In fact, on the simplex, the president represents $\hat{\alpha}$, for (s)he defines the moral narrative. The parliament (the people) is represented by high $\hat{\gamma}$: it represents the average will of the people. Finally, the law is in $\hat{\delta}$: the rational solutions that are laid down firmly (as deterministically as possible).

I can immediately deduce that, if the people are divided, the president loses power because he can no longer represent all the people. Therefore, the government (the parliament) and the law must gain power. Assuming the law is slowly changing, the parliament must take up the slack, and will resort to repressive sanctions to restore order. Rousseau describes in some detail the effects of population increase which can be easily tracked on the simplex. Suppose we start at the point halfway up the $\bar{\Gamma}\Delta$ edge. Recall that because of an arbitrary scaling, I can move any point in the 3D space to this point, so the analysis holds for any starting position. An increase in population necessarily entails a decrease in $\hat{\gamma}$ and $\hat{\delta}$, as diversity increases (so more people are doing and feeling things differently). Given that $\hat{\alpha} = 0$, only $\hat{\gamma}$ can be easily modified to account for this shift. It would need to be increased, and would also lead to more repressive sanctions.

However, an increase in what Rousseau calls “the relative power of the state” (Rousseau, 1762, p.104), which I have associated with the relative power of the president as the people’s emotional representative (so $\hat{\alpha}$ is increased), would also be able to restore order. What this means is that, if people were “forced” to cooperate more, they would increase their $\hat{\alpha}$, and become more reliant on each other, trust each other more, and use more restitutive sanctions. This would account for the move off the simplex caused by the increase in population. I have visualized this in Figure 18 by looking only at the mechanical vs. organic solidarity dimensions.⁴¹ Here, the increase in population forces the social order off the simplex, but order may be restored by increasing repressive sanctions or by increasing social cohesion.

Rousseau makes clear that these different factors occur at different levels, from individual, to corporate (e.g. parliament) to the state (everyone and the body politic). The institutional definition of the management of uncertainty will locate the individual on the simplex in each context. There may be tensions between various elements at that level, but the individual mind attempts to match *across all levels*. That is, an individual’s way of handling uncertainty becomes somewhat matched those around him by the free energy principle. At the same time, his actions influence government in such a way as to align it better with his views. Essentially, each individual, and each association (corporation) is participating in a big tug-of-war across the simplex. Some institutional bodies have well defined (e.g. constitutionally defined) power, while others are left as the sum of the powers of each individual. A well ordered state will have a better match between the people, government and the law, or at least create a body politic

⁴¹see Figure 8 for the location of these axes on the simplex - the actual axes, not the projections shown in Figure 2.

Figure 18: Changes in social order caused by increases in population size.

with just the right tension to establish an equilibrium somewhere on the simplex.

Rousseau was convinced that voluntary submission to the “sovereign” (to the law, for example) would allow the people to have positive and republican freedom, but not give up too much of their negative freedom (see Section 3.7). This requires each individual to define his self such that it is aligned with the group. This self is defined by looking in the mirror that is society (Cooley, 1902; Smith, 1759), attaining consciousness only through impressions on others. If this self is adopted by an individual, he is aligned with the group, and will feel like acting freely of his own will in exactly the way prescribed by the group. Thus, he gains positive and republican freedom, and loses only little negative freedom. He gives up natural rights for civil rights and for legitimacy (see Section 7.9 and Ridgeway’s legitimacy in Figure 2). As explored in Section 3.7, if the society “forces” its citizens to behave in a certain way (by restricting their negative freedoms), then an individual may cease to care about what is restricted to him, regaining negative freedom. Thus, the person stopped from crossing a dangerous bridge (that they do not know is dangerous) will feel their negative freedom is curtailed. However, if they stop even *wanting* to cross the bridge, they recover some of that individual negative freedom (Locke, 1690).⁴²

This is summed up nicely in the following: “*When a paralytic resolves to run, and when a fit man resolves not to move, both stay where they are*” (Rousseau, 1762, p.101). In the first case, negative freedom (free choice to run) is deadlocked with a lack of positive freedom (inability to run). In the second case, positive freedom (ability to run) is counteracted with a lack of negative freedom (resolve not to run). The non-runner in the second case simply *chooses*, through an act of will, not to run. He no longer cares about running, and so has complete positive freedom (he can run) and negative freedom (he could run if he wanted to). A similar example is reviewed in (Locke, 1690, p.149-150), where a man can stand still if he wills it, but if he “has not the power to remove himself,” he is not at liberty, just like the man falling from the cliff, who is not at liberty not to move.

⁴² Actually this is quoted in (Rousseau, 1762) in the introduction on page 36, and there is an example in (Locke, 1690, p.143) about a bridge but I can’t find the one where he is prevented from crossing the bridge.

6.1 Political Science

There has been a shift in political sociology away from uni-dimensional or categorical constructs of class (e.g. based on manual vs non-manual labor (Alford, 1963) and from occupation labels (Erikson and Goldthorpe, 2010)), to a two-dimensional view with one axis representing economic political views (from liberal to conservative) and social political views (from left to right) (Evans and Carl, 2018; Hiel et al., 2019). The social dimension differentiates between socio-cultural beliefs ranging from traditional to open and autonomous, and can be associated with the mechanical solidarity axis in Figure 8. When projected onto the simplex, I obtain the G bisector in Figure 2, shown as social left vs. right in Figure 19. The second dimension represents the economic differences that are based around more (left) or less (right) state control. This corresponds to organic solidarity in Figure 8, and thus to the G axis on the simplex in Figure 2. As found by Bienenstock et al. (1990), increased homogeneity and density tends to increase conservative views about a number of items (race, feminism, free speech and morality), and further to increase polarization between those with more liberal views and those with more conservative views. Associating homogeneity with increased $\hat{\gamma}$, this corresponds to movement along the G axis with increasing $\hat{\gamma}$, leading to a space of more socially and economically conservative views.

Figure 19: Dimensions of social and economic political beliefs shown as dashed lines on the simplex as projections of mechanical and organic solidarity, respectively, from Figure 7. Also shown major political parties ●=Republican USA; ■=Democrat USA; ◆=libertarian USA; ★=communist; ★=Despotism; and a key shift from communist to despotism rule.

The measure of density used by Bienenstock et al. (1990) is the “mean intensity or strength of ties joining alters” (Marsden, 1987, p.124), where “alter” is defined by those connected to a survey participant according to the “important matters” question of “who would you discuss important matters with?” Thus, this measure of density corresponds to the $\widehat{A\Gamma}$ with density=0 at $\hat{\alpha} = 0$ and density=1 at $\hat{\gamma} = 0$, and corresponds to increasing $\hat{\alpha}$, but *decreasing* $\hat{\gamma}$ (homogeneity). Thus, increasing density actually carries with it a *decrease* in homogeneity, which may help explain why some of the data in Bienenstock et al. (1990) do not confirm the hypothesis (although most do). For example, Figure 3c in (Bienenstock et al., 1990) is flagged as an anomaly (there is no

increase in polarization about race between males and females as density increases). This could be explainable because race often has external (visible) characteristics that imply it does not vary with the “trust” ($\hat{\alpha}$) dimension at all (or as much), but rather only (or more) with the homogeneity ($\hat{\gamma}$) dimension. Thus, any density effects can take place without any change in the $\hat{\alpha}$ dimension, which is what would cause polarization about race.

Facciani (2020) also investigated the effects of political heterogeneity and found that respondents would assign higher evaluation and potency to their ingroup (people with the same political stance), and would also view their ingroup as being more likely to engage in “helping” behaviours, while their outgroup is more likely to engage in “bullying” behaviours. Further, these effects were amplified (to the point of extremism) in homogeneous networks. That is, ingroup (outgroup) evaluation and potency were *further* increased (decreased) in homogeneous networks. Facciani (2020) also found evidence for homogeneity increasing extremism, but only amongst republicans. However, Facciani (2020) makes a number of hypotheses for which evidence could not be found, but I believe this may be because his basic modeling assumption that self-affirmation directly increases or inflates self-sentiments, and, and that self-uncertainty directly decreases or deflates self-sentiments, may be too simplistic. On (Facciani, 2020, p.2), for example, the connection between self-affirmation and an *increase* in self-sentiments (and between self-uncertainty and a decrease) is made explicit. In fact, self-affirmation is a form of uncertainty *reduction* as it *confirms* an individual’s sentiment. When self-uncertainty is decreased, connotative meanings obtain more salience and the *differences* in sentiments are therefore highlighted more strongly (not necessarily evaluation and potency). When self-uncertainty is primed, then denotative meanings become more salient, in which case it is even less clear how in- and out-group sentiments will be affected, and will likely depend on other unmodeled denotative factors.

Hiel et al. (2019) has shown negative correlations between right-wing attitudes (primarily social) and emotional intelligence. The measures used to quantify the different constructs (emotional intelligence and the two political view axes) are justified, but their exact interpretation on the simplex I am presenting is not altogether clear. I believe that emotional intelligence scales correspond to the D axis in Figure 2, while the other two are as outlined above. Thus, increases in emotional intelligence correspond to increases in **both** $\hat{\gamma}$ and $\hat{\alpha}$, and would correspond to a shift in both the social political viewpoint towards the left, and the economic political viewpoint towards more conservative views, although this shift is much smaller. However, given the geometry of the simplex (the illiberal axis is somewhere between the rationality and dependence axes), I expect to see large increases along the social political dimension, resulting in the stronger correlations as reported in (Hiel et al., 2019). However, it may well be the case that individuals (Belgian undergraduates) with more emotional intelligence are also more liberal minded, and so have higher $\hat{\alpha}$ to begin with. This would mean the study sample is biased towards less right-wing attitudes for more emotionally intelligent participants. As they control for education, this may be sufficient to reduce this bias, although it is not clear it will be eliminated. Thus, the results of a strong negative correlation between EI and social right-wing attitudes, and a weak negative correlation between EI and economic right-wing attitudes, are both consistent with the dynamics I have presented.

The arguments in Anderson (2017) center around the fact that the three freedoms exist to different degrees and that there are fundamental trade-offs between them. Further, the trade-offs may be different in different intersecting social structures. Thus, decreasing negative freedom at the state level (so people are subject to more rules constricting their actions) may induce

large increases in positive freedom (more options are available) and republican freedom (action is not arbitrarily specified within the set of possibilities) on the individual level. These increases may overall benefit individuals in terms of the total amount of freedom granted. Similar ideas can be traced back to Rousseau (1762).

Anderson’s (Anderson, 2017) three dimensions of freedom shed light on the two dimensional structure of economic and social left and right adopted by Evans and Carl (2018). In this view, which is developed from Rokeach (1973) (see Section 7.12), political beliefs can be characterized along a *social* dimension of conservative vs. liberal, and an *economic* dimension of right vs. left. In this case, economic right implies what traditionally be labeled as economic liberalism (free market *laissez-faire*), but we adopt the perspective that the center of this dimension is close to that employed by politically somewhat neutral countries, like Canada, with the Conservative party of Canada slightly more economically conservative than the Liberal party of Canada. In fact, Evans et al. (1996) explicitly interprets Rokeach’s dimension of *personal freedom* with libertarianism (what I have been calling “liberal”), contrasted with authoritarianism, and thus to Bales’ Forward (F) axis, and from there to *BayesAct*’s $\hat{\alpha}$ dimension. Evans et al. (1996) then interprets Rokeach’s dimension of *equality* with socialist vs. *laissez-faire*, which I can connect to Bales’ P/N dimension and from there to *BayesAct*’s $\hat{\gamma}$ dimension (which high values of $\hat{\gamma}$ corresponding to more *laissez-faire* economics).

In fact, the economic axis of Rokeach is my G axis, and the social axis of Rokeach is my A axis. If I displace the intersection of this point into the center of the simplex in Figure 2, then I get a diagram as follows, showing the two axes now centered, and the simplex as a dotted triangle. If this is the case, then we can also place the extreme positions of monarchism, communism, and Robinson Crusoe, and note the freedom’s of Anderson to produce Figure 20.

The two dimensions appear at least as far back as Hobbes, in which the *form* and *powers* of the state are considered different questions (Hobbes, 1651). Hobbes is considering invalid environments, and proposing state control (the “mortal God”) either through assembly or through monarchy. This axis (Hobbes’ *form*) corresponds to my $\hat{A}\hat{\Gamma}$, and since Hobbes felt monarchy was best (because non-replaceable assembly becomes a dictatorship), I have placed him in the far corner of high $\hat{\gamma}$, corresponding to pure and absolute monarchy. However, this state requires perfect homogeneity amongst all the major interests of all the citizens. Citizens in the “in group” have their action sets decided by some arbitrary will (so they lack republican freedom), but in exchange get great positive and negative freedom. This, in the words of Friedman (1962), requires one to exclude “madmen and children” as citizens, but leaves the definition of “madmen” open to interpretation. Therefore, to achieve this state, one might be tempted to find a group of people with recognizable physical attributes, and put them into this category. Women were such a group for many centuries. Hobbes’ other axis, that of *power*, corresponds to the illiberality in Figure 20, or to the A axis in Figure 2.

Descartes pursues a highly denotative solution and proposes that human will is deterministic as well, leading to a lack of free will (Descartes, 1637). This lack of free will corresponds with the lack of positive freedom in the highly valid world Descartes was proposing. His argument was that, in order to preserve the (proved) existence of God, human thought must not be causing the world to evolve, but rather human thought is also evolving deterministically, but completely separately, and the perceived explanations can be anything you want because of the complete negative and republican freedom, but they are nevertheless completely determined and therefore all positive freedom is removed.

Figure 20: Dimensions of social and economic differences in 2D with important German (black), American (red) political parties shown, and extreme “utopic” societies at the apices of the simplex (shown dashed) are shown in green.

Political parties, in their discourses, trade-off between the three freedoms and equalities. That is, they promise to reduce/increase taxes, reduce/increase services, or reduce/increase rules or law enforcement. The different parties represent groups of people who see the world in the same way: they use the same freedom/equality tradeoffs. This means political parties can be located on the simplex by analyzing their policies in terms of the three freedom/equality axes.

Let’s look at a few extreme examples of political systems *off* the simplex first. In the far back corner in Figure 20 (point labeled “freedom” and ■ in Figure 8) are true anarchists, whose state cannot exist because nothing would ever get done, or everyone would kill each other. In the top front corner (not shown explicitly in Figure 20, at the point labeled “equality” in Figure 8) are true communists, whose state can also not exist, because of mass confusion and disagreement about the definition of state. On the simplex lie workable social systems. Modern American Republicans, who are illiberal and economically conservative, would lie in a position of high mechanical solidarity, and low organic solidarity, or one of high republican equality, coupled with high negative freedom. They are shown as a red circle (● in Figure 20). As shown by Lührmann et al. (2020), the US Republican party has shifted a great deal *up* the

illiberality axis in the last 20 years. Modern American Democrats, and most Canadian parties, are somewhat near the center, but perhaps a bit higher on the organic solidarity and a bit lower on mechanical solidarity, or a bit more negative equality, a bit more republican freedom (■ in Figure 20).

Highly rationalistic states (with no positive freedom) were promoted throughout the 18th and 19th centuries (Kant, 1785; Marx, 1867), but are vulnerable to exploitation by despots, a feat accomplished in the 20th century to the detriment of a great number of people. Such despotic autocracies replace rationality with divinity, and strip citizens of republican equality, replacing it with positive freedom. This shift towards the economic right (less government) by a socially right wing, yet economically left wing communist party (★ in Figure 20) is possible with a despotic leader who, by re-defining what is meant by “rational” or “normal”, and by nihilating or marginalizing (de-humanizing) certain sub-groups, moves the group towards the *dependence* corner (★ in Figure 20).⁴³ These marginalized groups, labeled as non-human “barbarians” are revealed to be so because they are denying their own nature (which is, in fact, unbeknownst to them, the newly defined “normal”), and the result is a fascist state. Liberal political philosophy contains the same loophole, in fact, as everyone is free to act as they please, but republican freedom is only guaranteed by some democratic decision-making process that must exclude people incapable of making an informed choice in a certain way, e.g. “madmen and children” (Friedman, 1962). That is, the point at which decisions are made by a government attempting to balance a liberal view with other, more centrist views, must exclude those not able to tell the difference, but leaves open how to define “madman,” and has the same vulnerability to despotism (e.g. the despot can label some group as “madmen” or “savages” or “animals” in order to exclude them from decisions), as I will review in Section 9.7. It may be possible to avoid this using an automated method of finding a neutral decision making point that is independent of any political party.

Libertarians are big on positive and negative freedom, and are willing to sacrifice some republican freedom to get it. Being anti-government and secular, they require some kind of authority, but are happy to hand this over to big corporations (aka private governments (Anderson, 2017)). They are shown as a ◆ in Figure 20. The final US group of “progressives” (also labeled “woke left”) are those who place enormous significance on negative equality, ensuring the rights of all are guaranteed. They are fine with a heavy state, but are libertarian-like in their refusal to adhere to rationalistic views. They thus fall down in the far corner labeled “coherence” in Figure 20. Most social groups lie somewhere in the middle region of this space.

Movement across the simplex is possible, and may be caused by exogenous changes like natural disasters, or willful political, social or economic change, as in the despot example above. Another example is Durkheim’s study of the shift from mechanical to organic solidarity during the

⁴³The despot does this using a bait-and-switch trick. He (used *with* gender prejudice) first convinces the people that a rational way of doing things will bring everyone unlimited positive and republican freedom. He further convinces the people that *he* perfectly embodies this rational way (perhaps by divine right) and that he is perfectly aligned with the people’s ideology. He may easily do this by engaging with his targets in “theoretical squabbles that are quite irrelevant to their interests” (Berger and Luckmann, 1967, p.125) (e.g. abortion or gay marriage). Once he has the buy-in from the people (the alignment of their moral matrices with his, and the acceptance of him as ideological leader), he starts to move his own policies until he has removed sufficient republican freedom from the people, gaining enormous resources in the process. This is documented through history (Orwell, 1945). In fact, this trick can be used from anywhere on the illiberal side of the simplex in Figure 20, as groups in this region require a divine leader who can define things as he pleases. On what I have labeled the liberal side, this is more difficult, as individual identity takes precedence over divine meaning.

industrial revolution caused by social and economic changes (see Section 7.1). We can also consider the shifts caused by the events of 2020, as freedoms were massively changed very suddenly. Societies may react to these shifts by reverting to their group identities, likely causing increasing political polarization and other negative effects. Careful considerations of the multi-faceted problem of freedom and equality may be helpful for social progress.

The relationships between one- or two-dimensional views, and those of three-dimensional views (as *BayesAct* and Bales) are that the two-dimensional views are missing one of the dimensions. Even if they succeed in validly defining two orthogonal dimensions that lie on the simplex, they are still missing the third dimension, with the result being a somewhat confused analysis across the literature as many of the two-dimensional views are, in fact, using different sets of two dimensions. Attempts to characterize the three dimensions of uncertainty management using a two-dimensional plane run into difficulties because of this, and usually end up placing some vague constraints on the dimensions so that the resulting space is on the simplex, and is therefore a “ternary” diagram (Martin, 2009). I have noted some of these difficulties in Section 7.12, and here we see the same difficulties arising. Once the space is viewed in three dimensions, analysis can leave “flatland” and many of these difficulties may disappear.

6.2 Government, Sovereign and Law

In fact, the same analysis can be carried one step further if I allow multiple sub-sets of groups to exist, and I can even further make things interesting by also allowing three important subgroups that have substantial power: the government, the president, and the legislature. These three elements of society, in a free energy formulation as above, are each represented by some functional F . You can, Dear Reader, imagine three elements like this on a simplex, with elastic bands running between them. As one moves, the others react. Once these three powers are introduced, each is given a specific domain. The legislature has some control over $\hat{\delta}$, the president over $\hat{\alpha}$, and the government over $\hat{\gamma}$. This is simply because the legislature can only change the environment, not the people, the president can construct narratives that sway people’s morals, and the government can only do the will of the people.

Once this is set up, a giant tug-of-war occurs, with each element attempting to move in a certain direction. If the group is split, say into two parties, R and D, then the model θ_g becomes $\{\theta_{gD}, \theta_{gR}\}$, and these two each influence each of the others in certain ways. Thus, the sage ruler, can, if he wishes to move his state in a certain direction, compute the free energy of a configuration near where he currently sits, closer to his goal, and attempt to enact policy such that the various groups move in the direction required.

On new year’s day, 2021, the US Congress *overrode* a veto by the president on a military spending bill. The military spending bill is required to keep the military funded for another year. Many people depend on this. The congress in the US knows this and wants to approve this bill. As normal, checks and balances occur and the president vetoes the bill. Why does he do this? Because the bill contains certain provisions about social media limitations, and about base names in the US military. Both of these issues are basically emotional ones. He is objecting to a bill that requires a name to be changed because of his people’s emotional attachment to that name. This concern (his veto) should not be minimized. There are people that will find it very difficult to accept a name change, even at the cost of having their son or daughter stay overseas, or not get paid. Thus, the president espouses the emotional “core” of the people, and tries to

sway government in this direction. In the case of this override, the US government has firmly shifted the state towards the right, in my view, even though they are denying the principles of a large fraction of right-wing people.

Consider the problem of climate change. The primary problem with the human element in climate change is to get people to change their behaviours. The primary source of carbon is industry, followed by activities of private individuals. These two are linked of course, as the industries only exist for the individuals. However, government taxes on large corporations are seen by those large corporations instrumentally as incentives to move their business elsewhere, leading to an increase in unemployment. The only way this would work is if the world somehow all implemented these taxes, a very challenging problem. The other option is to change the activities of private individuals.

Interventions that change the activities of private individuals (what I have called “the social order”) can be changes to three things:

- the Social structure
- the Economic structure
- the Language or Narrative

So, let us imagine that you, Dear Reader, are interested in changing the social order. You may be doing this for all sorts of reasons I won’t probe you on, but you will have to change the three things listed above, which may be hard to coordinate. However, you can move in the right direction by changing just one piece of one of them. For example, you could change the economic structure by adding a tax, or you could change the social structure by making people more heterogenous (by moving them into cities) or more homogeneous (by segregating them). Both these laws would be enacted by the legislature, but supported by the government. Finally, you can change the narrative, which corresponds to changing the language. Examples are the lignite (coal) miners in Germany who felt offended by language calling fossil fuels “dirty”, in which case, a re-assignment of their identity (to e.g. “community steward” or “immigration mentor”) through direct government intervention (official titles) could be a potential solution. Consider again the manly construction worker who installs solar panels but would normally not consider buying a more efficient vehicle unless it was as “manly” as his truck. A narrative surrounding “manly” compact cars (e.g. in a popular TV show, or in a political speech), may be able to change this, and would fall to the president.

If you change only one thing, you move people away from the social order, and they will move back to (possibly, you hope) a new social order that more closely fits with your ideals in terms of climate change (note there may be other simultaneous changes being made by other departments in the government that will affect this in unpredictable ways). You can regain control of this movement by also changing one of the other two things, however the precise way in which these three are traded off by a given population with a given problem will be quite different. Therefore, one must choose a pair of these carefully and it might be easier to just change one at a time.

6.3 Neutral Point Decisions

BayesAct could be used as a tool for governments to assess the relative fairness of their decisions, including of decisions made by algorithm or AI (Stark and Hoey, 2020). Usually, government

attempts to handle a resource request, such as raising a tax, or tendering a contract, by evaluating how it will impact the freedoms of various groups of people, and trading this off against economic gains and losses. For example, a decrease in taxation may increase positive freedom for some individuals, while decreasing it for others, while increasing economic activity and decreasing government budgets at the same time. The savvy politician or political group will evaluate these freedom changes, and will apply some weighting function to make a final determination. Many current governments make use of algorithms to provide automated estimates of freedom changes and economic impacts that result from decisions (so-called *algorithmic government* (Engin and Treleaven, 2019)). They then apply a weighting function that may be heavily partisan. In contrast, *BayesAct* can be used to provide automated weighting of outcomes to ensure a neutral political choice. Thus, instead of changing the state to match the people (which is often not possible to do in a polarized state, and leads to some groups being marginalized, or worse, to a state vulnerable to a despotic ruler), it may be possible to change the people to match the state. Each person is likely to accept a neutral point decision, because the method I am proposing gives them two very good reasons to. First, the decision is the “will of the people” as it reflects their sentiments. Second, no one is better off than anyone else, and there are no free riders. Someone may be worse off, but only because they hold more extreme views.

Current global challenges facing all of human society, such as climate change and immigration, require a global institution that can enforce global laws aimed at reducing emissions, or moving people about. Such a global institution will face substantial challenges in finding solutions that somehow balance the needs of all member states, and departure of unsatisfied members will lead to a weakening of the institution. A neutral point solution still leaves each member state free to do as they please, but they will face increasingly stringent laws the further they deviate from the globally agreed upon prescriptions about globally relevant structures. What is interesting is that, in order to achieve a balance in a system of automatic government, one of two things can happen. First, the neutral point may be re-defined (shifted from the theoretically neutral spot) to better fit the current ethnography. Second, people may be educated in the ways of a minority group, in order to shift their median point back to the theoretically neutral spot.

Suppose a group that is split 50/50 on some issue is using a neutral point decision mechanism. Let’s call these two groups A and B, for simplicity. At the current 1-1 equilibrium, all are equally satisfied (or dissatisfied). Suppose now that a large fraction of *one* of these groups leaves (say half of the Bs emigrate to a more well suited country). Now, the balance is thrown off, and the current 2-1 split in favor of As leads to the neutral-point decisions appearing unfair to them, as it is too “B”-like given the current demographics. How can they deal with this? They can request a re-set of the neutral point, which can be considered given the current neutral point, and will naturally be found given the demographic shift.⁴⁴ That is, the automatic government itself will evaluate the neutral point reset, and find it equitable. However, this is also likely to lead to even fewer Bs, and could be catastrophic for the remainders. They can also import more Bs, possibly even recruiting some from the ranks of As through education. Finally, they can get rid of the Bs directly in a genocide, or more subtly through changes in fertility-related institutions. Some combination of these three effects can also be used. In any case, the remaining population is again at equilibrium with respect to the new or old neutral point, respectively.

I return to these ideas in Section 9.7.

⁴⁴A system of referenda and voting to implement this is needed, see Section 9.7.

7 Cultural Sociology

Since Pythagoras, the quest has been on for a denotative explanation of cultural structures and group processes. However, this has not led to fruition and still lies in a jumbled chaotic mess. The thing that was left out was the emotional or connotative component of reasoning, and if we look therein, we get a much simpler view. In this simpler view, the cultural structure resolves to the management of uncertainty in a deep Bayesian model. The basic model has two layers, one that contains information about the denotative state, and the other about the connotative state. As a “dual-process” type model, it combines affective and deliberative reasoning in a principled Bayesian way.

The tradeoffs between connotative and denotative meanings in reasoning link social psychological theorizing across many authors. The idea traces back at least to Durkheim’s mechanical vs. organic solidarity (Durkheim, 2014/1893), is also reflected in Lawler’s instrumental vs. relational commitments (Lawler et al., 2009) and in Bales’ dimensions (Bales, 1999). Individualistic groups in mechanical solidarity use instrumental commitments, and will require authority to control them and force them to obey social norms (through e.g. repressive penalties and enforcers). They are thus following “instrumental” commitments (Lawler et al., 2009), are acting in “mechanical solidarity” (Durkheim, 2014/1893), and must be more accepting of authority (Bales’ forward dimension (Bales, 1999)). In more diverse groups (such as those created originally in the industrial revolution with the amalgamation of agrarian people in cities supporting factories), social complexity pushes connotative reasoning to take over, and more collectivity develops in which “organic solidarity” (Durkheim, 2014/1893) and relational commitments (Lawler et al., 2009) are more salient, and groups are less accepting of authority (more of Bales’ backward dimension (Bales, 1999)). Such groups will self-regulate, but allow diversity in a population. Thus, when organic solidarity takes over, more collective groups in use relational commitments, shun authority, but obey social norms spontaneously because of trust (Fukuyama, 1995). Finally, I consider Lawler’s *normative* commitments (Lawler et al., 2009) to be my D axis, with high $\hat{\delta}$ values corresponding to fixed, normative behaviours encoded in e.g. laws.

Interestingly, I find that the three primary cultural sociological theories reflect reasoning threads that correspond to some degree with an increased dependence on one of the primary uncertainty factors (Lizardo and Strand, 2010). Therefore, I am proposing that cultural sociology research is politically biased. Many will agree with this statement without batting an eye, but many will not, so I will present a seafaring metaphor that will give a broad overview of the argument. More detailed analysis follows in the subsequent subsections.

Imagine a fleet of ships attempting to move in concert in some direction, surrounded by pirates. All ships have the same goal, but they don’t necessarily agree on how to get there. The ships are very vulnerable alone on the sea (because of the prevalence of pirates), and much safer if in a group. Some kind of structure needs to be set up to ensure that all ships remain together, i.e., that they can resolve their differences about how to deal with the denotative uncertainty presented by the weather. Traditional sociological methods (Weber, 1915) imagine one can work out how the weather works, and compute an optimal solution, which, since its optimal, will be the same for all the ships, leading to them moving in concert. The method is to find all the relevant factors and include them in a deterministic model. One could also see this as rafting all the ships together and corresponds with a high $\hat{\delta}$: everything is predetermined and predictable. This leads to a communistic political philosophy, as everyone’s negative freedoms are maximized

(although their positive freedoms are zero). Toolkit approaches, on the other hand, abandon this futile plan (as the weather is so complex), and instead come up with some simple “rules of thumb”, such as a flocking strategy (stay close to the mean position of the group, but not too close to any neighbour). These rules essentially remove the weather from the picture (decreasing $\hat{\delta}$). Now all that matters is the immediate: does my steering agree with yours? This is an increase in $\hat{\gamma}$, corresponding to an economically right-wing attitude. The overall pattern is the same. As the ant on the beach, the momentary actions are in direct reaction to the environment, but create the impression of an emergent order (Martin, 2010). However, this technique comes with a key flaw. The loss of grounding to the environment (reduction in $\hat{\delta}$) creates a meaninglessness to the whole operation - the goal is lost. A powerful leader, a Leviathan, is needed to be at the head of the group with a big red flag flying so all can go in the right direction. Perhaps multiple layers of such leaders will be required should the group get big enough. The boats have perfect negative and positive freedom, but are all restricted by the movement of the lead boat, so they lose all republican freedom.

The last method is the socially and economically liberal one, and corresponds to the cultural sociological *practice* theory (Bourdieu, 1990). In this method, each simply places additional ballast in its hold. Suppose there is a current flowing below the surface. These ballasted boats, if uncontrolled, will end up in the same place due to the current. Further, the motion of the boats in the same direction as the current *increases the current* as the water is carried along with the boats, entraining those behind.⁴⁵ Thus, all are guaranteed to stay together, to move to a common place, and to require no steering or sailing (almost effortless!). But wait! Of course, you, Dear Reader, has grounds to stamp your foot and declare in a huff that this common place to which we are heading now may be in *the opposite direction of the goal!*. Indeed, we have two choices. First, we can release our desire to achieve that goal, increasing our positive freedom (decreasing $\hat{\delta}$) and accepting a high $\hat{\alpha}$, a perfect lack of negative freedom. Second, each boat can maintain the original common goal, and steer *with respect to the current* in the direction of the goal. Suppose just one boat starts doing this, the current *created* by their slight motion slightly misaligned with the primary current will entrain some nearby boats perhaps, growing the effect. The end result *could be* a complete change of the current, reverting to the previous situation. However, this requires a *commitment* on behalf of all to follow this plan, such as to prevent reduction of republican freedom.

Thus, in practice theory, the direction to steer is provided by the overall social group (the current), but the individual action has some leeway with respect to that current, also forming the current at the same time. As in *BayesAct*, high $\hat{\alpha}$ gives a clear direction to go (with the current), but individual positive freedom allows each to steer with respect to that current how he sees fit. Again, it requires all to commit at the same time, so a level of trust is required to make this system work. Note that the motion *with* the current is imperceptible should no boat be attempting to steer. Crewmembers on the rebel boat steering away from the current will perceive themselves as moving sideways through a stationary fleet. However, the current they *create* will entrain other boats, possibly leading to a runaway effect. Otherwise, the other boats

⁴⁵I do hope the sailors amongst you, my readers, will forgive my complete lack of knowledge about sailing. This is not meant to be an accurate model of sailing, but rather only a fanciful metaphor for the third method of social coordination. You can, if you wish to make it even more fanciful, thus freeing yourself from your prejudices about sailing, then you can imagine the boats all being turned upside down, and having big rudders. The rudders allow for some vague steering using the wind, and this can be used to modify the direction of the whole group, but it takes time.

have only to extract tithes from the rebel boat as it goes by (for forcing them to counter-steer back into the current).

BayesAct may be summarized by looking to the seven concepts or devices that describe “culture” as outlined in Small et al. (2010), and re-iterated in (Vaisey and Valentino, 2018). These seven concepts have associated concepts in *BayesAct* as follows:

- **Values** correspond to the parameters for the management of uncertainty (coherence, dependence and validity), and the dynamics (changes to) these parameters as a result of interactions. This interpretation of values corresponds to *evaluative knowledge* (Patterson, 2014; Martin and Lembo, 2020), and so steps one level back from describing the “ends toward which behaviour is directed” (Small et al., 2010, p.14) to describing the ways in which those ends can be computed.
- **Frames** correspond to denotative beliefs (priors and posteriors). Frames are heavily used in artificial intelligence as, for example, statements about what conditions need to occur before a certain plan can be carried out (pre-conditions), or what the effects of actions will be (Poole and Mackworth, 2010). Frames may include *goals*, or more generally *utility functions* that encode preferences. These declarative elements of an agent’s value system are those the agent seeks to optimize over, but only within the limited vision afforded by the blinders of his connotative system, of his values. In fact, it may be the case that values are reified into objective reality in any institution as preferences of particular types of agents, and thus allows bootstrapping of the utility function onto the values. However, starting with a prior set of preferences can help the agent get going in the right direction.
- **Repertoires** correspond to denotative policies or strategies of action, referred to in artificial intelligence as *heuristics* and in cultural sociology as *toolkits* (Lizardo and Strand, 2010).
- **Narratives** correspond to denotative dynamics and histories, and can be used to establish trajectories that were followed in history or that may be followed in the future. Narratives are strengthened by the denotative coherence of the story being told, and are often constructed “after the fact” as an explanation that justifies certain behaviours (Mercier and Sperber, 2017; Haidt, 2001; Berger and Luckmann, 1967). After an alternation of self (e.g. a religious conversion, a change in relationship), negative narratives about the old self, may serve to bolster support, beyond what is provided by positive narratives, for the new (Berger and Luckmann, 1967, p.160).
- **Symbolic Boundaries** correspond to the somatic transform, and associate morality, culture, and identity to narratives.
- **Cultural Capital** correspond to connotative beliefs in general, and in particular over high-status elements of culture (so there is an ordering of capital). In *BayesAct* the restriction of cultural capital to only high-status elements is relaxed, and capital can be held over any connotative belief. In a criminal gang, for example, the capital may be the perceived and transmitted “bad-ass-ness” of a certain individual.
- **Institutions** are a somewhat broad umbrella, corresponding to both connotative and denotative “programs” in *BayesAct*, and so encompass all of the above in some sense. Institutions are “called up” from memory upon entering some institutional context, and imply a

values setting, prior beliefs (denotative and connotative) and a set of strategies/heuristics. All three conceptions of institutions mentioned by Small et al. (2010, p.19) are handled in *BayesAct*:

- formal rules of behavior (norms in *BayesAct* that are enforceable),
- norms of behavior that are not enforceable (toolkits/heuristics)
- taken-for-granted understandings that frame (connotative beliefs)

The tradeoffs I have been discussing above are implemented in *BayesAct* to explain a number of cognitive biases (Hoey et al., 2020), to relate to theories of emotion (MacKinnon and Hoey, 2020), and to explain relationships to the sociology of culture in this paper. *BayesAct* has the three parameters of uncertainty described above as either settable (in the case of sufficient prior information) or learnable for any specific group of people that is sufficiently large to gather statistics on.

In this section, I will relate the model presented in Section 2.2 with some elements of cultural sociology. I do not claim this to be an exhaustive review, but I hope to give some pointers to how these relationships may be made. My goal is to propose a unification of these elements, which I believe ties the management of uncertainty tightly to the anthropological and sociological literature. This exposition does rely on an understanding of the Bayesian model, however, so is not a stand-alone read. However, the reader who has justifiably skipped the more mathematical sections will be able to make some use of this section by studying the following associations from symbols to meanings. These are the axes and apices of the 3D space and simplex shown in Figure 2, which should also be studied by the reader who wishes to make the most sense of this section. These symbols will also be used in the text, so it may be worth keeping them handy, and therefore, dear Reader, I have put them in Figure 3 on a separate, printable, page.

7.1 Durkheim

Emile Durkheim (Durkheim, 2014/1893) had a conception of social structures which was grounded in studies of the movements of peoples and of the changes of sentiment during the industrial revolutions taking place in the late 19th century. A key idea Durkheim expresses in *The Division of Labour in Society* is that an agrarian monarchy shifted to an urban social contract under the influence of industrialization. As people moved to big cities and did repetitive factory jobs, diversity (heterogeneity) and opportunity (invalidity) were key new factors of uncertainty that changed how people grouped together, how they were solidary. This can be used to explain a number of observable phenomena, for example, people may start having too much validity at work, and not enough at home, so that these two spheres would move apart, and the work and home would become more separated. Durkheim characterized this solidarity in two ways: as *mechanical* and *organic*. The definition of these two solidarities is intimately tied up with the definition of the *collective consciousness*, defined on p.63 as

“The totality of beliefs and sentiments common to the average members of a society”

I can then infer that Durkheim is referring to the overall *shared* posterior belief (both connotative and denotative). With this in hand, I will start with Durkheim’s statement about the genesis of solidarity:

“Not only does mechanical solidarity generally bind men together less strongly than does organic solidarity, but, as we mount the scale of social evolution, it becomes increasingly loose.” (p.119-120)

This reflects the transition from groups bound together by denotative understanding of the world to groups bound by connotative meanings, and so indicates a movement along the $\widehat{AI}$ axis from purely mechanical solidarity at high $\hat{\delta}$ and $\hat{\gamma}$, transitioning to more organic solidarity along the $\widehat{\Delta A}$ axis, with high $\hat{\delta}$ and $\hat{\alpha}$. This also connects to Lawler’s “instrumental commitments” transitioning to “relational commitments” which have much the same character (easily replaceable) see Section 7.10.

Durkheim lists three conditions that modulate the strength of these social bonds. I associate these with three functional forms in *BayesAct* that arise in the combination of probability distributions. The first modulation condition reads:

“1. The relationship between the volume of the common consciousness and that of the individual consciousness. The social bonds are stonger the more completely the former overlaps with the latter.” (p.120)

This is the overlap between F_i and F_g (see Section 4.3), and essentially means that each individual, and the group as a whole (and the body politic) share the same posterior, both connotatively and denotatively. The bonds overall will be stronger the more they overlap. They overlap best when free energy is minimized, which is when everyone matches each other. This is easy to do for any group by simply increasing $\hat{\alpha}$ and decreasing the other two. The result, however, is a utopia of flower children, with most concern being on social graces, good feelings, and equality. Such a society would disintegrate because nothing would actually get done. Thus, a certain temperance is required, hence the construction of law, the president and constitution, and the government (the “body politic”).⁴⁶ These institutions are designed to pull the society in a direction towards higher $\hat{\delta}$ (the law) and higher $\hat{\gamma}$ (the government). So overall, this first condition is enforcing a state and body politic that both lie in some kind of equilibrium on the simplex.

Durkheim then decomposes collective consciousness into two components: intensity and determinateness, and shows how low intensity, indeterminate collective consciousnesses implies less mechanical solidarity (and thereby more organic solidarity, to be proved). On the other hand, high intensity, determinate collective onsciousness implies more mechnical solidarity. Thus, I associate mechanical solidarity with an increase in $\hat{\gamma}$ and $\hat{\delta}$. In both cases, similarity of individuals is increased. Thus, Durkheim’s second modulation condition reads:

“2. The average intensity of the states of collective consciousness. The relationship between the volume of the common and individual consciousness assumed to be equal, the degree of intensity has more effect upon the individual the more energy it possesses. If, on the other hand, that intensity is caused by only a feeble impetus, its capacity to steer the individual in a collective direction can only be feeble. Thus, the more easily he will be able to go his own way, and solidarity will be less strong.” (p.120)

⁴⁶I am using an American example throughout, but one could easy replace the President with the Queen of England, and remove the constitution, to recover something close to the Canadian case.

The intensity is thus the strength of the motivational force catalyzed by the collective, or the precision of the posterior distribution in $\mathbf{Y}$ given any value of $\hat{\alpha}$. It is thus represented in *BayesAct* with $\hat{\gamma}$.⁴⁷ The third modulation condition defines the determinateness:

“3. The degree of determinateness of these same states [of collective consciousness]. Indeed the more beliefs and practices are clear cut, the less room they allow for individual differences. They act as uniform molds in which we all cast, in a uniform fashion, our ideas and actions. Consensus is therefore as perfect as possible; every consciousness beats as one. Conversely, the more general and indeterminate the rules of conduct and thought, the more individual reflection must intervene in applying the rules to particular cases. But such reflective thinking cannot be aroused without disagreements breaking out.” (p.120)

I associate this component with high $\hat{\delta}$. That is, as the beliefs and practices become more clear-cut, agents become more and more similar, and mechanical solidarity is again increased. In a highly valid world, for example (high $\hat{\delta}$), agents will act exactly the same way as they are all following the same program (the optimal one for that domain denotatively). In invalid worlds (low $\hat{\delta}$), the “degree of determinateness” has less effect, leaving the intensity of the collective to take over.

Taken together, Durkheim’s statements about collective consciousness indicate that (1) sharing of understanding; (2) agreement on meanings; and (3) intensity of beliefs all contribute to how much solidarity one expects to find in a group. Once this collective consciousness exist, Durkheim’s definition of “solidarity” on p101 (Durkheim, 2014/1893) points to two primary, but opposing “forces”: one centrifugal (drives individuals away from collectivity) and the other centripetal (brings individuals together). The first such force links individuals to society “without any intermediary” (p.101), and agents relying on this force cannot depend on other parts of society with complementary skills or services, as the agent has no control over these elements. Individuals are maximally independent, but require everyone to be working on the same overall project in perfectly aligned worlds in order to make a functioning society. I can therefore draw a parallel between this force and the $\hat{\delta}$ and $\hat{\gamma}$ vertices in the simplex in Figure 2: high individuality, but requiring coordination to function properly. This type of “mechanical” solidarity was more common prior to the industrial revolutions, with small groups (e.g. families) living in agrarian communities in which each individual provided a complete service (e.g. farmer). The farmer was completely independent of all others in his overall function, but relied on others for some specific, instrumental, needs.

The centripetal force is provided by the link of individuals to society through their dependence on the “parts that go to constitute it” (p.101). This type of solidarity is a system of “different and special functions” (such as those created by the industrial revolutions), which are “united by definite relationships”. This directly calls for a connotative management system in which different “identities” are used to define these definite relationships. The key idea is that segmentary society (where people lived in small groups that were well defined) had a strong connotative collective consciousness defined by the denotative state through a strong somatic transform (so everything was the same, and everyone felt the same). Those in urban and factory settings were forced out of this strong connotative consciousness because of their (enforced) denotative

⁴⁷You can imagine this by picturing the $\hat{\alpha} = 0$ case. In this case $\mathbf{Y}$ is *completely determined* by $\mathbf{X}$.

diversity. They instead developed an “organic” solidarity based on relationships that could be directly modeled connotatively for reasons I have stressed above.

Prior to the industrial revolutions, agents living in mechanical solidarity had strong collective consciousness: everyone was exactly the same, and sentiments and beliefs were very strong and well defined, and everyone agreed on the meanings of things. Thus, they existed on the simplex at the point where the A axis crosses the $\widehat{\Gamma\Delta}$ edge. However, there was less possibility of also having shared normative understandings (moral contracts), and so there had to be an authority. The individual had no mechanism to handle deviance from the collective consciousness, and so this needed to be handled by a higher power. During industrialization, people moved together in cities, density increased as did diversity: people with different sentiments and beliefs came together and so $\hat{\delta}$ and $\hat{\gamma}$ necessarily decreased. Since this decrease then made things less predictable, people needed some way of predicting (with some degree of certainty or trust) what others would do, implying a strengthening of affect control, and so $\hat{\alpha}$ must have increased. People needed to do this in order to be able to collaborate at all. Each person needs to “trust” that the other person is going to work towards a common goal, in order to give him the motivation to continue attempting to align somewhere in this mess of large sentiment distributions. A central controller implemented as enforceable rules (a Leviathan) becomes more difficult to implement because of diversity: there is no one set of rules that would work for all. Further, since $\hat{\delta}$ decreased, agents with larger $\hat{\gamma}$ would have more difficulty in finding cooperative solutions with other agents, as they would more often be “wrong” in their sentiments. Thus, $\hat{\gamma}$ is further forced to decrease, leading agents more firmly towards solutions that are highly cooperative and transitive (with high $\hat{\alpha}$).

Durkheim talks about “*disagreements breaking out*”(p120) if the degree of determinateness is reduced, and then about cooperation leading to “*conflicts and quarrels*”(p167) if duties and rights are not defined. Thus, we can see that fights will occur if either $\hat{\delta}$ or $\hat{\gamma}$ is decreased, since the center of this space denotes a state of pure freedom in which the body politic does not exist (anarchy). Thus, assuming $\hat{\gamma}$ cannot be changed (the city’s heterogeneity is fixed), then new legal structures are needed, corresponding to an increase in $\hat{\delta}$. However, new legal structures (new forms of contracts, for example), are burdensome in a competitive market, and players with less contractual obligations can operate in an economically equally secure way without the overhead. If they do this, they are opting to be *less* secure economically, but placed the burden on the connotative system, implying a system of good graces and appropriate behaviour, for example. This implies that higher $\hat{\alpha}$ is important, as it enables more efficient transactions, and better stability in the face of external forces, as they can more quickly adapt to a changing environment. This is why, incidentally, businessmen play golf and drink scotch together. It establishes these relational bonds that are so important in a diverse and global society.

Thus, we can associate the emergence of organic solidarity with the increasing dependence on $\hat{\alpha}$ as $\hat{\delta}$ and $\hat{\gamma}$ were “forced” to decrease during the industrial revolution. In fact, they may not have been “forced” as much as they were optimized to be that way by a capitalist ruling class, who benefited from the transition enormously. Further, due to the increased efficiency of transactions when $\hat{\alpha}$ is increased, the decrease on $\hat{\delta}$ could be counteracted to a certain degree, leading to a more horizontal transition from high $\hat{\gamma}$ to high $\hat{\alpha}$. This is shown in Figure 21.

Summing up the argument:

“For organic solidarity to exist, it is not enough for there to be a system of organs

Figure 21: Simplex on $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ showing the location of mechanical and organic solidarity, as well as the transition that occurred during the industrial revolution in the 19th century (shown as a red arrow). Locations of repressive and retributive law also indicated.

necessary to one another that feel their solidarity in a general way. The manner in which they should cooperate, if not on every kind of occasion that they meet, at least in the most common circumstances, must be predetermined. Otherwise, a fresh struggle would be required each time in order to bring them into a state of equilibrium with one another, for the conditions for this equilibrium can only be found only by a process of trial and error, in the course of which each person treats the other as an opponent as much as an auxiliary. Such conflicts would therefore break out continually, and in consequence solidarity would be hardly more than virtual, and the mutual obligations would have to be negotiated anew in their entirety for each individual case.” (p.285)

In less developed society everyone is the same and the collective consciousness is strong. Motivation to act comes from the fact that each individual knows who he is and who everyone else is. Any deviance must be dealt with through repressive sanctions, rather than restitutive ones. Thus, there are fixed social programs created by external forces which themselves are created by the similarity of individuals. Everyone is the same. Everyone agrees that outcome A is of value, while outcome B is not. Everyone strives for outcome A. Agents in this type of society do not have strong predictive models, and so tend to think of shallow, broad (more rational) solutions, rather than deep and narrow ones (“flashlight” algorithm again, see Section 4.2).

Repressive law implies that the transgressor will be punished, while restitutive law implies that the victim will be rewarded (often by the transgressor). As shown in Figure 21, repressive law can exist only when $\hat{\alpha} = 0$, for in this case the individual has *no rights*. That is, the leader can impose any arbitrary penalty on him. Further, restitutive law will *not* work when $\hat{\alpha} = 0$, because in this case there is no motivation to cooperate, trust is zero, so everything becomes valueless. Consider money as a trust that is implied between the holder and the government. With $\hat{\alpha} = 0$, this trust is gone, so the money is valueless. Therefore, restitutive penalties will

not work. However, restitutive sanctions will work (and repressive sanctions will not work) when $\hat{\gamma} = 0$, because this implies $\hat{\alpha}$ is high, so individuals have rights that will be violated by repressive sanctions, but will value restitution at least as much as rights.

Now, the most important point made in Durkheim's book is that liberty is increased as solidarity becomes more organic. People living in mechanical solidarity are *less* individual, but they don't know it.

“His liberty is only apparent; his personality is borrowed”(p.314-315)

As $\hat{\alpha}$ is small, members *think* they have more freedom as they appear to be considering more action choices, but yet since they are all the same ($\hat{\gamma}$ and $\hat{\delta}$ are high), they in fact do not. The liberty in Durkheim's quote is a liberty of acting, which requires a structured consensus of individual meanings. Everyone is the same, but can do what they want. However, since all members are the same, they will all do the same. The liberty of organic solidarity members, on the other hand, comes from the ability to be different from one's neighbours, and to have the ability to change at any time. The society is flexible and intense. It is a liberty of being, which requires a structure at the level of the social action. Everyone is different, but must get along. The increased flexibility that comes with decreased $\hat{\delta}$ and $\hat{\gamma}$ requires a mechanism for individuals to cooperate (through increased $\hat{\alpha}$).

The bottom line is that as society progresses, organic solidarity becomes possible, which leads to more individual liberty of being, more efficient and long-term strategic cooperative (social) thinking, but requires more equality in external forces. Games being played by members must have more equal payoffs. This is the very basis of socialist thinking.

“[the idea of higher societies] is to inject an even greater equity into our social relationships, in order to ensure the free development of all those forces that are socially useful.”(p.302)

7.2 Lévi-Strauss

In his classic work, Lévi-Strauss (1967) examines the dual processes that exist within “totemic” cultures, where a set of symbols (e.g. animals) are used to denote both relations with other humans, and relations with nature

“As I have already suggested, ideas and beliefs of the 'totemic' type [...] constitute codes making it possible to ensure, in the form of conceptual systems, the convertibility of messages appertaining to each level.” (p.90)

The poles of these “levels” are then characterized as being of two types, based around culture (men's relations to each other) and nature, and the mediation between the two being the “distinctive function of the totemic operator.” (Lévi-Strauss, 1967, p.91) Finally, that the transcendence of the opposition between the two is the primary function of totems. That is, the totem is a narrative used to explain or account for natural events (e.g. what not to eat), but makes it easier to motivate action because of its emotional potential. I can associate the narratives defined by totems then as a representation of the connotative meanings of culture, while the

denotative meanings relate to the natural (uncontrollable) world, and the somatic transform as the operator that transcends the two.

The structuralist ideas of Lévi-Strauss (1967) also point to three dimensions of variation.

“The process of artistic creation therefore consists in trying to communicate either with the *model*, or with the *materials*, or with the future *user* as the case may be [...]” (p.27, emphasis in original)

And that

“All forms of art allow all three aspects and they are only distinguished from one another by the relative proportion of each” (p.27)

I can associate the *model* with $\hat{\alpha}$, as a focus on the model leads to abstract art, disconnected from reality, art for art’s sake, and would be dominated by relational, affective meanings. *Materials* are associated with $\hat{\gamma}$, in which the immediate relationship to the world is of primary importance, and it “internalizes the occasion” (immediate) but externalizes the “execution and purpose”. Finally, a focus on the future user is associated with $\hat{\delta}$, and becomes apparent in the applied arts. These are crafts that serve a purpose (the execution and purpose are internalized), but that have less dynamic meaning

“[...] the supernatural beings which it delights in representing have a reality which is timeless and independent of circumstances”

Now I’m sure I could go on for some time about the connections with this vast literature, but instead I will hypothesise something and leave art critics to decide if it fits their experience. I propose that art, is, in fact, a third solidarity dimension taken up by the anthropologists. A projection of this third solidarity dimension onto the simplex is labeled as D in Figure 2, but the dimension itself originates at the origin, and grows as $\hat{\alpha}$ and $\hat{\gamma}$ grow, as shown in Figure 8.

7.3 Bales

Robert Freed Bales (Bales, 1999) describes a three dimensional structure that has a close relationship to the simplex I have presented in this paper. In fact, Bales describes a two-dimensional surface embedded in a three dimensional space, as I will now show.

Bales describes a “cube” defined on a three dimensional space of “values” with axes as follows:

label=P : Friendliness, values on “equality, democratic participation in decision making”.

label=N : Unfriendliness, values on “self-protection, self-interest first, self-sufficiency”.

label=F : Acceptance of authority, values on “conservative, established ‘correct’ ways of doing things”.

label=B : Rejection of authority, values on “change to new procedures, rejection of conformity”.

label=U : Dominance, values on “individual financial success, personal prominence and power”.

label=D : Submissiveness, values on “giving up personal needs and desires, passivity”.

These axes can be mapped to the dimensions of uncertainty I have explored in Section 2, by noting that acceptance of authority (F) occurs with low values of $\hat{\alpha}$, while rejection (B) would be high values. Similarly, unfriendliness (N) occurs with high values of $\hat{\gamma}$, while friendliness (P) with low values. Finally, dominance (U) occurs with high values of $\hat{\delta}$, while submissiveness with low values.

However, the two three-dimensional spaces do not quite match up as the Bales axes are defined on the space of real numbers: if we map it to a Cartesian space, then one half of each dimension would be negative values (e.g. B is the negative of F, D is the negative of U, and N is the negative of P), and the *BayesAct* space is in the space of real *non-negative* numbers (the parameters $\hat{\alpha}, \hat{\delta}$, and $\hat{\gamma}$, being measures of variance, cannot be negative). This means that very *small* values of $\hat{\alpha}$, for example, corresponds to large values on the F dimension in the Bales space, which is the negative of the B dimension.

To map the *BayesAct* space into the Bales space, I introduce a simple sigmoid “squashing” function as follows, mapping a parameter $\Psi \in \{\hat{\alpha}, \hat{\gamma}, \hat{\delta}\}$, where each of $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ has a domain of $[0, \infty]$, to a Bales dimension $L \in \{b, p, u\}$, where each of b, p and u has a domain of $[-\infty, \infty]$:

$$\Psi = \frac{1}{1 + e^{-L}} \quad (11)$$

Or, alternatively, that the Bales’ dimensions are a *logit* function of the *BayesAct* dimensions:

$$L = -\log \left(\frac{(1 - \Psi)}{\Psi} \right) \quad (12)$$

One example of this function is shown in Figure 22. Note that there is a scale factor I am not accounting for in each dimension, meaning that the y axis (the actual values on the Bales’ dimensions) are somewhat arbitrarily scaled in the figures I present.

Figure 22: Logit function mapping *BayesAct* space (Ψ) to Bales space (L).

This mapping is shown in Figure 23 by simply plotting the two sets of axes in the same coordinate frame. The actual mapping between the two spaces is difficult to show, but one pictures that

Figure 23: (a) The simplex in $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ space is mapped to a simplex in Bales' space, but with apices at infinity. Thus, I have gathered the points along each edge the simplex at angles of 30° and 60° in the Bales' space, and added a small number 0.1 to each of the dimensions that have value 0 for that edge. The result is a hexagon, as shown in the figure. One has to picture that the $0, 0, 0$ point in $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ space maps to $(-\infty, -\infty, -\infty)$ in this space, and thus the hexagon is a “cut-out” in the middle of the infinite simplex in Bales' space. (b) the original Bales drawing (fig 2.9), rotated and flipped so it lines up with (a), showing the same regions numbered as in (a), see text.

the $(0.5, 0.5, 0.5)$ point in the *BayesAct* space maps to the origin of the Bales space. The origin of the *BayesAct* space maps to $-(\infty, -\infty, -\infty)$ in the Bales space.

The origin of the Bales space maps to the point $(-\infty, -\infty, -\infty)$ in the $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ space, while the apices of the $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ space lie at points $(\infty, -\infty, -\infty)$, $(-\infty, \infty, -\infty)$, and $(-\infty, -\infty, \infty)$ in the Bales space, making the resulting simplex somewhat difficult to draw. Instead, I can simply “cut off” the planar surface so defined, and thus draw it in the Bales space as a hexagon as shown in Figure 23, mapping the points along the edges of the simplex at angles of 30° and 60° for each of the three dimensions in *BayesAct* space $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$.

I can then map the different value types given by Bales, using his numbering into the three dimensional parameter space $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$ as follows (also shown geometrically in Figure 23). First, I convert the Bales space to *BayesAct* space by assigning values of $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$ to the numbered groups shown in Figure 23(b) that correspond roughly to their position on that graph. In the following table, the symbols have the following meaning: $\uparrow$ means large (0.6), and $\uparrow\uparrow$ means huge (0.9), $\cdot$ means negligible or mid-range (0.5), $\downarrow$ means small (0.4) and $\downarrow\downarrow$ means tiny (0.1). I have mapped these using Equation 12 to values in the Bales’ space, and then plotted them in Figure 23 as the corners of the hexagon formed by “cutting off the corners” of the simplex, as explained in Figure 23.⁴⁸

Bales region	cultural group	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\delta}$
1	most effective teamwork core	$\downarrow$	$\uparrow$	$\cdot$
2	liberal teamwork side	$\cdot$	$\uparrow\uparrow$	$\downarrow$
3	conservative teamwork side	$\downarrow\downarrow$	$\cdot$	$\cdot$
4	group-centered wing	$\uparrow$	$\uparrow\uparrow$	$\downarrow$
5	authority-centered wing	$\downarrow\downarrow$	$\downarrow$	$\uparrow$
6	swing area	$\cdot$	$\cdot$	$\cdot$
7	libertarian fringe	$\uparrow\uparrow$	$\uparrow$	$\downarrow$
8	individualistic fringe	$\downarrow$	$\downarrow\downarrow$	$\uparrow$
9	anti-group opposition	$\cdot$	$\downarrow\downarrow$	$\cdot$
10	anti-authority opposition	$\uparrow\uparrow$	$\cdot$	$\cdot$
11	Radical opposition core	$\uparrow$	$\downarrow$	$\cdot$

This agrees with Bales’ Figure 2.9 (p61) (Bales, 1999), reproduced in Figure 23(b). Note there is an somewhat arbitrary relative scale between the axes, as previously mentioned.

This picture gives us some evidence for my discussion of the three axes and their relation to political values. We can see that the “liberal teamwork side” for example, works best in a fairly homogeneous environment, but with some transitivity, and a moderate amount of rationality. The “libertarian fringe” on the other hand, works best in a reasonably diverse environment, with some transitivity, and not much rationality. I return to this in Section 9.

This shows that, although the Bales space defines “values” as lying in all three dimensions, these are from studies of work teams that are not necessarily coherent. Bales describes the optimal coherent group as all lying the $P - U - F$ direction (region 1), which maps to reasonably high values of $\hat{\gamma}$, mid-range values of $\hat{\delta}$, and small values of $\hat{\alpha}$. However, he also points out that different groups may operate in different areas of this space. Although my analysis shows that the “optimal” parameter settings for group coherence lie on a simplex in *BayesAct* space,

⁴⁸These points would also be a hexagon in *BayesAct* space.

there may be in fact large variation around that simplex (perpendicular to it) across any given working group. This will typically occur due to incoherence in the group about the settings of these parameters, and may cause the group to operate sub-optimally. That is, some group may have members who do not “fit” with the team.

Interestingly, Bales indicates a correlation between forward and more conservative political beliefs, and between backward and more liberal political beliefs (Bales, 1999). Thus, from a computational sociological point of view, we are led to the suggestion that the political spectrum of beliefs is defined by uncertainty and ambiguity management techniques: conservatives tend to use undersocialized models (higher $\hat{\gamma}$ or $\hat{\delta}$), while liberals use oversocialized models (higher $\hat{\alpha}$). This corresponds to the illberality axis in Figure 19.

7.4 White

Combine the disparate radiation spectrum of species presented by White (1992) as leaning on molecular and ecological contingent determinants that annex the elemental dimensionality structures adumbrated in a triplex of orthogonality-infused embeddings and self-similar processes with the valuations of required biophysical essences of flux through obfuscated indeterminateness. Put another way, White (1992) describes the same structure of three dimensions of uncertainty as I have put forward in this paper. However, his description takes on many different forms, some of which I will describe in this section, leaving the remainder for future work.

White is essentially proposing that the distribution of individual’s parameter settings (generative models) within a group will lead to certain structures which reinforce these very same settings, some more strongly than others. Three extreme such distributions are considered by White (1992), which he defines as *disciplines*. I therefore propose a *BayesAct*-based definition of *discipline* that potentially can help disambiguate some of the important bases of White’s book: the set of parameter values in a given group, and their relative positioning and correlations on the simplex of Figure 2. White adheres to the Comtian tradition of analogy between molecular biology and social structures, and refers to a *species* as a setting of the three “embedding ratios,” which corresponds to a location on the Shepard plane and on my simplex.

White clearly defines three dimensions of uncertainty: differentiation, dependence and involution (White, 1992), which are also referred to as contingency, ambiguity and ambage.⁴⁹ These dimensions are taken up by Martin (2009) (see Section 7.5). I can make an association between differentiation and $\hat{\delta}$, dependence and $\hat{\gamma}$ and involution and $\hat{\alpha}$ based on my analysis of Bales (1999) in Section 7.3. White (White, 1992, p.29) explicitly points to a relationship between his *value ordering* of *quality* and his first *species of discipline*, the *interface*. He points further to a symmetric relationship between his valuation ordering of *purity* and the *arena* discipline, and of *prestige* in the *council* discipline. Finally, he relates *quality* to Bales’ instrumentalism (F/B), which maps to my $\hat{\alpha}$, *purity* to Bales’ P/N axis, which maps to my $\hat{\gamma}$, and finally *prestige* to

⁴⁹The relationship between the two sets of three dimensions is as follows. Contingency is the amount of denotative precision, so maps directly to differentiation. Dependence maps to the inverse of ambiguity (high dependence is low ambiguity), while involution maps to the inverse of ambage. That these are parameters of dispersion in a probabilistic model is pointed to by the ranges being from $0 \rightarrow 1 \rightarrow \text{high}$, see (White, 1992, Fig.6-1). Figure 1-2 shows these concepts as well, but the contingency axis is not flipped while the other two are, making this diagram somewhat difficult to interpret geometrically. These dimensions are also referred to as “temperatures” (White, 1992, p.271), which indeed they are in the *BayesAct* model: the parameters are the “temperature” parameters in a Boltzmann distribution (see Section 4).

Bales' U/D axis, which maps to my $\hat{\delta}$. I can then, by studying (White, 1992, Fig.2-1) or (White, 2009, Fig.3.1), associate differentiation with $\hat{\delta}$, dependence with $\hat{\gamma}$ and involution with $\hat{\alpha}$. The triangles in White correspond to the simplex, but the edges of White's triangle in his Figure 2-1 (reproduced as Figure 24(b) in this paper) are labeled in a somewhat counter-intuitive way as they correspond to the perpendicular bisector in Figure 2 defined by the quantity in question. Therefore, in Figure 24(a), the dashed red lines corresponding to my A, D and G bisectors are labeled with the discipline that occurs along that line, towards the far end from the label. That is, the interface occurs at high values of $\hat{\alpha}$, where the $P(Y)$ is relatively undispersed, and well shared amongst the group. This is shown in Figure 24(c), where interfaces are shown as green dashed lines at different values of involution ($\hat{\alpha}$), so the main tradeoffs along these lines are between $\hat{\gamma}$ and $\hat{\delta}$. Comparing this to (White, 1992, Fig.2-1), reproduced in Figure 24(b), we can label these green dashed lines as lines of constant involution ratio, while the two edges (in this case $\widehat{\Delta A}$ and $\widehat{A\Gamma}$), are labeled *differentiation ratio* and *dependence ratio* by White (1992). The same logic can be applied to the other two disciplines, leading to the magenta and cyan dashed lines in Figure 24(c). Note the rotation between Figure 24(b) and (c).

Figure 24: (a) White's space drawn as equivalent to the *BayesAct* space; (b) (White, 1992, Fig2.1); (c) disciplines shown on the simplex, showing the relationship to (White, 1992, Fig2.1).

To see more precisely how this relation works, I examine the most precise model specification I could find, in (White, 1992, Appendix.2). This clearly states how the *interface* discipline involves a balance of *differentiation* and *dependence*, and referring roughly to my G axis (or any line on the simplex parallel to $\widehat{\Gamma\Delta}$), and defined by the *quality intensity*, and thus the amount of *involution* in the group, corresponding to my $\hat{\alpha}$. If $\hat{\alpha}$ is large (if the “*quality intensity of commitment*” (White, 1992, p.331) is high), then an interface results, in which a group can stably “lock onto” a specific shared parameter setting of $\hat{\alpha}$ (amount of involution) that is reinforced by this commitment. This stability then leaves any diversity in the group to differences in dependence and differentiation as noted by (White, 1992), and as shown as green dashed lines in Figure 24(c).

The two other disciplines are similarly described relative to the dimension values at or near the two other apices of the simplex. The *arena* discipline is defined by a balance between involution and differentiation, such that a strong *purity valuation* (corresponding to a high $\hat{\gamma}$ in my model) will tend to produce groups that strongly share values of $\hat{\gamma}$, but then vary across values of $\hat{\alpha}$ and

$\hat{\delta}$ (roughly the D axis, shown as magenta lines in Figure 24(c)). Finally, the *council* discipline is defined by a balance between involution and dependence, where high values of $\hat{\delta}$ lead to groups with strongly shared denotative models, but with varying interpretations connotatively (roughly the A axis, shown as cyan lines in Figure 24(c)).

There are far too many other aspects of White’s ideas that I cannot cover here, but I will make note of a few. First, the idea of a *story* is one that corresponds to a denotative set of (probabilistic) constraints in the *BayesAct* model, other toolkit model, or some other cognitive bias, possibly a social *norm*. White also focusses heavily on *identity* as a shared social construct (see pp.205-209 in (White, 1992)) that arises from irregularities and mismatches in the social embedding. “*Persistence through crisis is where identity emerges*” (White, 1992, p.208). From a *BayesAct* perspective, this seems to be indicative of the following sequence of events. First, a breakdown in the social order occurs, and the dispersion of the posterior in the denotative state. In turn, this sparks an increase in focus on the connotative state, and a seeking of new denotative state to account for the increased variance. Should this new denotative state include some action for the agent (which it must), White refers to this as “*getting action*”. This new denotative state can take the form of a new identity for a person, which may lead to a new social structure. This social structure then leads to denotative (normative) constraints, which serve to “*block*” (in the sense of restrict or constrain) action.

White’s concepts of “*getting*” and “*blocking*” action map to exploration and exploitation in reinforcement learning (see Section 8). To “*get*” action is to decrease connotative dispersion (increase $\hat{\alpha}$), which leads to inequality (White, 1992, p.257,p.293), and also to more focus on connotative beliefs, and therefore increased density of the posterior over non-rational options. Referring to this as “*exploration*”, this creates new denotative state, and “*gets*” action (in the sense of “*creates*” action, meaning that new actions are discovered). However, those denotative structures that are the result of all this “*getting*” of action become normative constraints, thus “*blocking*” action, known in reinforcement learning as “*exploitation*”. These constraints are also referred to as “*heuristics*” or “*toolkits*” (see Section 7.9).

As a simple example, consider again the grass-eating agent of Section 2.1. This agent can distinguish grass from not-grass, but cannot distinguish tiger-grass from not-tiger-grass. The breakdowns in denotative predictions that are a result of this inability (the agent gets chased and nearly eaten, say) lead to the immediate creation of new (powerless and negative) connotative state, and new denotative identity of the grass-eater as *prey* to correspond with that new denotative state, and new action of *flee* that corresponds to what *prey* does when threatened. *Control* is a mixed denotative and connotative control, and *decoupling* is the dis-association of the two through a weakening of the somatic transform. The denotative control is connected to decision-theoretic control (White, 1992, p.235), particularly the dynamic programming of (Bellman, 1957), which is one method of finding policies in a Markov decision process. In a social situation, this type of inflexible control is criticized, essentially for ignoring the blocking and “*getting*” of action that the connotative state enables. White’s focus on self-similarity and the “*fractal*” nature of the world is possibly a bit of a fanciful and unnecessary adornment of multi-layer (deep) processing, as are his multiple forays into connections to cellular processes. Nevertheless, the multi-scale nature of intelligence does connect a massive range of scales from cellular and to astronomical, and so these apparent flights of fancy may have meaning waiting to be discovered (Friston, 2019). The many other potential connections between *BayesAct*, Bales (1999) and White (2009, 1992) are certainly fruitful grounds for further exploration.

7.5 Martin

Martin (Martin, 2009) analyzes much sociological reasoning behind these problems, and presents multiple instances of a ternary space that I associate directly with the simplex in Figure 2. I have noted the key elements of three of these ternary diagrams in the figure at appropriate locations to denote this association. I will review them here briefly.

Relationships (Martin, 2009, p69) Preferential attachment ($\hat{\delta}$ apex) is the primary source of social structure in a completely denotative world. In this case, the only way to ensure there is some kind of structure is to associate with someone more powerful than oneself who defines rationality. Along the $\widehat{A\Gamma}$ axis, there is no individual structure, but more influence of affective factors expressed as reciprocity or transitivity. This is what Durkheim calls low individuality, although acknowledges that it involves high *denotative* individuality (diversity). The $\widehat{\Gamma\Delta}$ axis is labeled as “no group structure” as it is something that must be imposed from outside (by a Leviathan). This also falls high on the B axis of Bales (1999), corresponding to groups that respect authority.

Stratification (Martin, 2009, p101) In this figure, Martin associates the D axis in Figure 2 with stratification. This arises in much the same way as how preferential attachment arises. Exchange structures are built from transitive relationships, while homophily is built through strong shared meanings (high $\hat{\gamma}$). The G axis is then the “homophily” dimension (I have previously labeled this a “homogeneity” dimension, which is the same thing), while the A axis is the “impulse to spread” (favoring larger groups) dimension.

Structures (Martin, 2009, p223) Here, the structures that are expected are the multihub (for high $\hat{\alpha}$), the web (forming cliques) for high $\hat{\gamma}$, and the pyramid for high $\hat{\delta}$. In highly valid environments, for example, individuality becomes key, but all follow the same denotative policies. There is more “exclusivity” along the $\widehat{\Delta A}$ axis as there is more differential structure. When $\hat{\gamma}$ is high, the “web” structure has tightly connected individuals, but these can only be so big before cognitive resources are exhausted, forming cliques.⁵⁰ Within such cliques, the web structure ensures no exclusivity. When $\hat{\delta}$ is high, then the structure resolves to a pyramid (stratification and authority), while when $\hat{\alpha}$ is high, the multi-hub structure can emerge due to the transitive relationships.

Martin further points to some of these concepts being based around ideas presented by White (1992). I also have found a correspondence with White (1981), and this leads to some interesting perspectives on economic markets, but, dear Reader, I will beg your forgiveness and will leave that to present in a further paper.

7.6 Values and Interests

The *BayesAct* model considers reward or *utility* to be defined as a rank-order set of preferences over a set of outcomes. Utility, combined with a probabilistic model of the same outcomes, yields a decision theoretically rational policy of action. However, these policies are notoriously difficult to compute, particularly in situations where the outcomes are dependent on the actions of other intelligent agents (in which case, the decision theoretic solution “*blows up*”). Nevertheless, there is something to be done with this utility function, and that is to *justify* the actions that were

⁵⁰Etymology of “pair” comes from Latin meaning “equal”.

taken. Looking backwards through time, an agent may be able to identify that action a produced outcome o , and thereby transfer the preference rank of o onto a . This type of inferred action is then inferred as something that *should* be done, should o have high utility (Martin and Lembo, 2020). Should o have low utility, then a would be considered something that *should not* be done. These are referred to by Patterson (2014) as *espoused values*, and are ones reported on e.g. surveys.

On the other hand, what people actually *would do* in real situations often differs from what they report that they should do (Martin and Lembo, 2020). This is also known as a difference between values in use (would do) and espoused values (should do) (Patterson, 2014). The reasons for action in traditional cognitive models are often predicated on the creation of norms, sometimes including some component of emotion or sentiment, which are most easily added to the reward function (Fehr and Schmidt, 1999). However, the explosion of propositional structure that must be added to a denotative model to account for norms across people and contexts is enormous, and thus these manipulations are limited to somewhat specific situations, or situations with many control variables.⁵¹

Nevertheless, people are known to act in specific ways according to a set of values that are less precisely defined, but that act as “bright lights” (Hitlin, 2008) cited in (Martin and Lembo, 2020)), they “*still contribute something to action in a perhaps somewhat obscure way*” (Martin and Lembo, 2020, p.57). I propose to lift this obscurity with the the somatic transform in *BayesAct*, which defines a distribution over actions or policies (and thus futures) for any denotative state or observation. It gives this distribution immediately without any prior knowledge of the immediate situation, although it may be based on all sorts of history with this person or the context recalled from memory. This distribution is used to filter futures that are examined as possibilities, only exposing those that are affectively reasonable (that will cause the agent to better fit its econiche). Any choices are thus made over options that are connotatively reasonable (that affirm the individual’s position in the interaction or broader culture).

Norms in *BayesAct* can be learned, or can emerge as “encasings” of connotative values in denotative structure (e.g. labels, this concept from (Lawler et al., 2009)). That is, since these connotative structures are exposing the action space, the outcomes will be those expected given the context and culture, and will be following establishing a pattern in memory. These patterns, if often repeated, become *reified* and thereby “encased” in language. However, these norms have a uni-directionality which is their downfall. Although it is easy to go from label $\rightarrow$ encasing (norm) $\rightarrow$ connotative value (feeling), as used in *justification*, it is hard to go the other way, as used in policy search. If asked, our boy with the flashlight will proudly tell you how he knew just where the target was going to be, and can certainly produce the physics of the situation that would convince you, dear Reader, beyond a shadow of a doubt, that his actions were well reasoned. Even though his report may convince you that he could predict the physics of the mid-range, mid-sized target, there may be many explanations for his success, including a closer, smaller, target, or a farther, larger, target. Therefore, the boy would have difficulty actually making a prediction, and may rely as usual on a vague feeling that the targets would be more to the left, and go poking around there.

I argue in this paper that *BayesAct* unifies the *would* and the *should* by associating the current

⁵¹ Attempts to characterize these norms as properties of an individual follow the model of (Schwartz, 1994), which proposes three primary dimensions that can be mapped roughly to those of Bales (1999) and to my $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$ dimensions, however they are implemented denotatively by (Schwartz, 1994), see Section 7.12.

usage of the word *values* with denotative state, and the word *interest*, which corresponds to the somatic transform and the connotative state. Therefore, I refer to *values* in this paper as being the same as *reward* or *utility*. The degree of *interest* is computed from the posterior over the action distribution in *BayesAct* (see Section 4), which relates any denotative information about the scene with any connotative information available (including prior information).

The two examples of folk proverbs from Martin and Lembo (2020, p.57), “He who hesitates is lost” and “Look before you leap”, are examples of denotative encasings of connotative meanings of being “daring” and being “cautious”, respectively, and the relevance of these are dependent on connotative identities (e.g. a “mother” is “cautious” whereas a “teenager” is “daring”). There are many situations that caution could be applied to, so the denotative proverb is generally applicable. However, the denotative acts of leaping or looking are only valid in situations of low or high risk of some negative outcome $\mathbf{x}'$, say losing resources, respectively. After the fact, the actor can evaluate whether $\mathbf{x}'$ has occurred or not, and then recover the meaning of his action. If he acted, and $\mathbf{x}'$ didn’t happen, he was “daring”, but if he didn’t act, and $\mathbf{x}'$ did happen, he was “cautious”. The opposite cases (he didn’t act and $\mathbf{x}'$ didn’t happen, or he did act and $\mathbf{x}'$ did happen) are more difficult to handle, and may require re-identifications (he should have been more daring, or it was someone else’s fault, respectively). On the other hand, applying one of these denotative proverbs in practice involves actually *evaluating* the risk, something that requires much more computational resources because all possible futures have to be examined.

“[...] values can well be used to *justify* actions taken, but, contrary to the optimism of Parson’s original account, it is rather difficult to know how to deduce from them very specific instructions in any particular situation.” (Martin and Lembo, 2020, p.56-57)

The philosophy of science (PoS) also invests in understanding the nature of reward and value, as these have substantial impact on the direction science is taking at any time. In this literature, “values” in the most common usage of the term, or as they say, “à la Marx,” refers to preferences in a decision-theoretic sense, while “interests” correspond to the connotative meanings of things for a specific individual (Martin and Lembo, 2020). These are referred to in PoS as epistemic values (rewarding information seeking, exploratory behaviours), and non-epistemic values (rewarding exploitative security through similarity and trust). Epistemic values are rewards in the decision-theoretic sense (as fulfillment of preference). Non-epistemic values are successful relationships with others, fairness between populations, or similarity with others, for example (Fish and Stark, 2020; Elliott and McKaughan, 2014). Weisberg (2007) describes three forms of non-epistemic model building in science, which further support this claim: Galilean idealization ($\hat{\delta}$ - the model that works the best), minimalist idealization ($\hat{\alpha}$ - the most aesthetic or elegant model that fits with our interests) and multiple-models idealization ($\hat{\gamma}$ - pick the model that fits best, or average).

I can therefore relate non-epistemic values to model choices (parameter settings in $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$ space), while epistemic values are defined explicitly in any model as reward if using a decision-theoretic approach, or expected states if using an active inference approach. In *BayesAct*, these two values are traded off as (1) reduction of deflection in $\mathbf{Y}$; and (2) optimisation of expected reward in $\mathbf{X}$. For the same reason that people who are self-affirmed will be more exploratory because their denotative distributions are more dispersed (see Section 7.13), people pursuing non-epistemic values more strongly (less dispersed distribution in $\mathbf{Y}$ and $\pi(\mathbf{Y})$ or $\mathbf{f}_b$) will be more exploratory in terms of epistemic values (more dispersed distribution in $\mathbf{X}$ and $\pi(\mathbf{X})$ or

a). However, this is a double-edged sword, because more exploration in $\mathbf{X}$ is only within the increased constraints placed by the more precise distribution in $\mathbf{Y}$. In the limit, “exploration” in $\mathbf{X}$ becomes pure exploitation in $\mathbf{Y}$ (exploitation for the sake of non-epistemic values). At the other extreme, a dispersed distribution in $\mathbf{Y}$ shifts focus to epistemic values and reward, leading to pure “exploitation” of $\mathbf{X}$, which is pure exploration over non-epistemic values. In order for this to work, however, the agent will require a more *valid* environment (one in which risk can be more accurately estimated). As a process, a novel, highly invalid situation in $\mathbf{X}$ will shift focus to $\mathbf{Y}$ allowing the agent more flexibility to find alternatives in a , White’s “getting action” (see Section 7.4).

7.7 Moreno

Moreno studied groups using sociometry and found emotional meanings to be important as well (Moreno, 1934).

“Our problem is the social expansiveness of man and his transmission of emotion” (Moreno, 1934, p.162)

By studying the social structures in sociograms of rapprochement (“liking”), he came to propose three “laws” as follows (framed here in terms of *BayesAct*) (Moreno, 1947):

- The *sociogenetic law* states that societies and groups move from mechanical (instrumental) to organic (relational). This agrees with Durkheim’s transition, and Lawler’s theory of social commitments (see Section 7.10).
- The *sociodynamic law* states that the dynamics of Y are shared. This means that groups operate on the relational dimension ($\widehat{A\Gamma}$ edge).
- The *social gravitation law* is homophily and preferential attachment, implying that groups also operate along the intransitivity dimension (the $\widehat{\Gamma\Delta}$ edge).

Combining the last two laws, we would expect to see structures somewhere in the middle, pulled towards transitivity (group) and towards rationality/homophily at the same time with counteracting forces. These forces depend on the status of the individual relationships along the transition from mechanical to organic (along the D axis, the sociogenetic law). These laws were studied further on triadic sociograms by Feld and Elmore (1982), who found inequality to increase with the number of double or non-choices (along the $\widehat{\Delta A}$ axis), thus associating equality with the G axis in Figure 2.

7.8 Osgood/Heise/MacKinnon

Osgood studied cultural meanings across cultures, making remarkable findings on the nature of shared affect (Osgood et al., 1975). Heise (Heise, 2007) then combined these affective meanings with a control law, added dynamics and denotative state, and formulated affect control theory (ACT - see a more complete introduction in Section 4). Heise started with the control principle of Powers (1973), which states something very reminiscent of the free energy principle: that people try to minimize incongruencies by controlling their perceptions. He proposed these

incongruencies could be applied to the affective sentiment space of Osgood (E,P,A), both in and out of context, and combined this with a denotative definition of the situation (Mead, 1934) and added affective dynamics in a subject-verb-object grammar of interaction (Gollob, 1974). That is, the model proposes that a social situation is first appraised denotatively (identities are assigned to the involved agents), and then the affective dynamics create an impetus to act in a way that minimizes social incongruencies. Re-identifications can also be made denotatively (e.g. a “doctor” can be re-identified as a “nurse” after observations of certain affective behaviours). ACT is formulated propositionally in (MacKinnon, 1994), in which two fundamental principles are proposed. First, *inextricability* refers to the inability of the human mind to process denotative content without associating connotative meaning. This is similar to Weberian concepts of values and interests as “*inseparable*” (Martin and Lembo, 2020, p.78). Second, *complementarity* states that the two types of meanings are both required for social cognition. Martin and Lembo (2020) also point to interests and values being different, but both being a necessity (they are worthless alone). These issues are explored more in depth in relation to *BayesAct* and the concepts presented here in (MacKinnon and Hoey, 2020; Hoey et al., 2020).

Heise made attempts to unify the E,P,A space with that of Bales (Heise, 2013), based on the discussion in (Bales, 1999, p.140-143), in which Bales maps his 3D space to that of Wundt (Wundt, 1896) by associating his P/N with Wundt’s pleasant/unpleasant, and his U/D and F/B to a rotation of Wundt’s Tension and Arousal dimensions. Heise correspondingly maps Bales’ P/N dimension maps onto the evaluation (E) dimension, and Bales’ U/D and F/B map to a rotation of the power/activity (P/A) dimensions.⁵² However, I believe both are mistaken in attempting to unify E,P,A dimensions with Bales’ dimensions using a rotation. In fact, Bales even closes his attempt by saying that “The fit is problematic” (Bales, 1999, p.141). The fit is problematic because Bales’ space is a space of model *parameters*, not a space of emotional meanings. That is, Wundt, Osgood, Heise and Russell and Mehrabian (1977) are all discussing the same three dimensional space that is an explanatory subspace in which to characterise feelings or sentiments about things.⁵³ On the other hand, Bales is discussing a parameter space in which the processual nature of an interaction is guided by positioning in his 3D space. The two spaces are *not* unifiable with a simple rotation. I explore Bales’ space in more detail in Section 7.3, and show a potentially more parsimonious unification of the parameter space in *BayesAct* with Bales’ space. I also relate an anecdote in the discussion about David Heise’s reaction to this novel unification in Section 9.

7.9 Traditional, Toolkit, and Practice

Cultural anthropology has long sought to connect individual behaviour with group practice. Such accounts are usually highly context dependent, and attempts to generalize require exhaustive research and analysis across cultures (D’Andrade, 1984). Such context dependence, however, is the bread and butter for sociologists and anthropologists, and are more fully representational of real contexts than are those examined in laboratory settings and surveys. D’Andrade posits four levels of meaning, starting from *representational*, in which the world is represented, corresponding to the denotative relationship between observations and inferred hidden states. Moving

⁵²More precisely, F maps to P+A-, B maps to P-A+, U maps to P+A+ and D maps to P-A-.

⁵³Although I don’t see power in Wundt’s space, but rather a dimension of “tension” which is probably a measure of dispersion or uncertainty, often considered a fourth dimension of affective meaning (Fontaine et al., 2007).

to *constructive*, D’Andrade describes the creation of cultural entities, which I associate with connotative meanings in *BayesAct*. Next, *directive* meanings are directive of action, reflected in *BayesAct* as the posterior distribution over policies, also as the *width* of the flashlight beam in the flashlight allegory. Finally, *evocative* meanings evoke feelings, represented in *BayesAct* as the posterior over the connotative space (the size of the targets). D’Andrade points to two major intrinsic motivational systems: 1) direct personal reward; and 2) reward because of attachment to a particular set of values. Ridgeway echoes the same distinction (Ridgeway, 2019).

While the model in Ridgeway (to be examined in more detail below) is purely denotative, I have argued in this paper that this only gets us part way to a solution, as it ignores the emotional overtones pervading all behaviour and context (or treats them denotatively). The modeling of connotation is much rarer in the literature, but cultural sociologists make a distinction between heuristic cognitive biases (in so-called “toolkit” theories) and deeply ingrained patterns of behaviour (in so-called “practice” theories) (Lizardo and Strand, 2010). Toolkit approaches have found success in explaining social structures (Martin, 2009), and are hypothesised to arise from the scaffolding of the environmental and social structure. The scaffolding is so complex that humans learn heuristics and tricks to get by, but the tricks are “defined” by the scaffolding, since they are created in order to handle exactly it. Thus, from a distance, the social structures may look complex, but in reality each individual is following a set of simple rules, like Simon’s “ant on a beach” (Martin, 2010). Practice theory has been notoriously more difficult to apply, primarily because of a lack of operationalization (Li et al., 2009). *BayesAct* provides such a scaffolding by associating connotative meanings with all elements in an interaction, including with agent actions, thus providing a rough (heuristic) guide over policies. The social intelligence provided by connotative consistency is shared by agents in a community, and motivates them to want to do things according to the same practice or “habitus”, which encodes the “way we do things” (Bourdieu, 1990; Ambrasat et al., 2016). This shared practice is an approximation built to handle and alleviate the computational complexity of the social world, and guides an agent towards socially acceptable choices of behavior that can ensure more globally optimal solutions to social dilemmas.

In *BayesAct*, behaviours explained by toolkit and practice theories are reflections of the same underlying joint denotative and connotative dynamics (connotative dynamics, somatic transform, and denotative dynamics). Traditional sociological theories (Weber, 1915) are usually solely based on denotative meanings and logical relationships and as such would lie at the $\hat{\delta}$ apex. Practice theories require social cohesion and thus high $\hat{\alpha}$, while toolkit theories that include some emotional component (e.g. (Ridgeway, 2019)), need at the least a denotative model of the dispersions modeled by $\hat{\gamma}$. I can then associate the differences between these theories as lying along the A axis in Figure 2. Practice theories are highly context independent (low $\hat{\delta}, \hat{\gamma}$) and highly relational (high $\hat{\alpha}$), while toolkit theories are more rational (high $\hat{\delta}$) and more homophilic (assume agents copy one another) (high $\hat{\gamma}$).

In *status characteristics and expectation states theory*, Ridgeway (2019) proposes an essentially “toolkit” approach where

Status hierarchies and rankings, as a social form, result from a cultural schema, a blueprint, of learned and shared rules for organizing social relations in order to manage the tensions and coordination problems posed by contexts of cooperative goal interdependence.(Ridgeway, 2019, p.151).

The normative rules so proposed are those that represent the status hierarchy, and as such would contain propositions such as

nurses defer to doctors

meaning that, should you be a nurse facing a doctor, then your culturally normative action is to defer to the doctor. The normatively salient action is also the one rewarded. In *BayesAct*, such normative behaviour is not encoded in a reward function directly, instead making its way there through the attention mechanism of the connotative state. Affective dynamics are known to make such normative-like productions, and so focus attention on those parts of the denotative state that include deference, for example. This then defines the preferred states, which also map directly to reward, see Section 4.3. Our nurse would now *only perceive* those actions he could take (denotatively) that would also be deferential connotatively due to the presence of a person with the identity *doctor*. Thus, actions such as *request something from* and even *elude* would be part of his denotative deliberations about what to do, while actions such as *command* or *make love to* would be out of the question, so to speak. Surprisingly, the behaviours at the opposite end in affect control theory (as far away from “defer to” as possible) are *save* and *cure*, with *shoot* coming in third place.⁵⁴

Ridgeway describes three biases that help explain structures, however, and these biases, if encoded in *BayesAct* parameters, give rise to affective dynamics that reflect the biases. A *status bias* describes the “full range of evaluative biases about the relative competence and status-worthiness of different actors that are unleashed by diffuse status characteristics when they become salient in a social relational context”(p.121). I make a correspondence between the strength of these evaluative biases and the somatic transform precision, $\hat{\gamma}$. Ridgeway also refers to this as *esteem* (Ridgeway, 2019, p.79). Higher precision somatic transforms give rise to biases originating in e.g. power differentials between assigned identities. These identities can arise from status characteristics and may form stereotypes for example. The second of Ridgeway’s biases is a *legitimacy bias*, which gives actors with certain identities license to behave in certain ways. I therefore make a correspondence between this bias and the connotative dynamics precision ($\hat{\alpha}$). Ridgeway also refers to this as *trust* (Ridgeway, 2019, p.79), corresponding to the associations made in Section 6.1. Higher precision connotative dynamics ensures trust and implies such licensing. Finally, Ridgeway’s *associational preference* denotes a hierarchical ranking of actors based on competence, which I associate with the denotative dynamics precision. More valid environments (or more expertise in an agent) entrain more preferential attachment and creates, supports, or enhances inequality based on status characteristics. It is also known that higher competence groups tend to view with some moral suspicion, corresponding to their slightly higher positioning on the $\widehat{\Delta A}$ axis towards the $\hat{\delta}$ apex, relative to the group average or to the positioning of the lower status group members.

Ridgeway also discusses the tension between these three biases in some detail, noting that dominance hierarchies require perceptions of perceived competence (denotative), as well as status beliefs (connotative) about this same competence of other group members. This leads to a fundamental “tension between leaders and followers in a goal-interdependent group.”(Ridgeway, 2019, p.44). As groups move beyond dyads, the necessity of having individual competence (ability to solve the denotative problem) becomes inextricably coupled with a need to have shared beliefs about status (connotative beliefs about who should lead) which may be based on gathered evidence in a current interaction (e.g. behavioural or appearance-based cues about dominance).

⁵⁴Based on the Georgia 2014 data (Smith-Lovin et al., 2016).

Such status beliefs propagate cultural stereotypes as they are connotatively encoded. That is, groups can only adopt a more denotatively efficient solution in which the leader is followed by first establishing motivation in each agent to follow that leader.

In expectation states theory, the dynamics of status follows a reflexive feedback loop (Tilly, 1998). First, diffuse status characteristics are assigned based on observable status cues, such as gender, race, or physical artifacts. “...people jockey for position in the micro-dynamics of status through objects they display”(Ridgeway, 2019, p.113). Status of objects (e.g. a fancy phone) has been studied with an affect control theory framing in (Shank and Lulham, 2016). Second, these diffuse status characteristics modify expectations for performance in ways that are self-fulfilling. Agents with initially assigned high status are able to define the direction the group takes in ways that are most conducive to maintaining the status hierarchy so defined. This status may then be challenged or found lacking in evidence by the group. Alternatively, low-status actors may be able to rise in status should they display some assertiveness, but not too much as this will be seen as disingenuous or deceptive. For similar reasons, leaders need to display competence, but also fairness and interest in a shared group identity with their followers (Ridgeway, 2019, p.149)(Steffens et al., 2020). In ACT and *BayesAct*, these processes are explicitly modeled using identity dynamics, so that inputs X give rise to beliefs about identity in Y , which then “guide” the search for action (flashlight algorithm) and naturally re-inforce the hierarchy by forcing lower status groups to realize the denotative expectations of higher status groups who control resources. The lower status group is then forced to accept the connotative meanings defined by the higher status group, acceptance of which involves a re-definition of their position (as one of deference). Changing status for an individual involves re-defining identity for that individual within the group, a process requiring group support that may not be forthcoming.

Ridgeway works from evidence describing status formation in groups based on hierarchies as “normative systems for managing cooperative interdependence to achieve shared goals”(Ridgeway, 2019, p.47). Ridgeway attributes the functionality of such normative hierarchies in groups to the fact that “interpersonal status is experienced psychologically as insecure rather than settled” (Ridgeway, 2019, p.60), and states that “status beliefs are also a tool to assuage psychological uncertainty”(Ridgeway, 2019, p.67). Finally, that

“When people’s sense of control over a social situation is threatened (for instance by the psychological uncertainty of status), they have an increased preference for structure and hierarchy. Access to a shared set of clear, relatively stable status beliefs that rank social attributes and behaviors provides a form of such structure and hierarchy that meets individual psychological needs as well as collective goals.”(Ridgeway, 2019, p.67).

These notions are firmly in line with the model I am proposing here. Status beliefs based in *social identities, attributes and behaviours* (Ridgeway, 2019, p.67) are linked to psychological uncertainty about relative status positions, modeled as dispersion in the connotative state in *BayesAct*. As control is threatened, however (as this dispersion is increased), then people will want to rely more on the denotative explanation of the situation. As I have argued, a denotative solution to a group problem can only be handled through an authoritative “forcing” of structure and hierarchy.

However, in order to ensure such status norms are adopted, members adopting low-status positions will need to be convinced to do so. Although this behaviour is explained using emotional

		B		
		defer	cooperate	defect
A	defer	(2,2)	(3,4)	(0,4)
	cooperate	(4,3)	(2,2)	(0,3)
	defect	(4,0)	(3,0)	(1,1)

(a)

		B		
		defer	cooperate	defect
A	defer	(3,3)	(4,4)	(1,4)
	cooperate	(4,4)	(2,2)	(0,3)
	defect	(4,1)	(3,0)	(1,1)

(b)

Figure 25: Normative games equivalent to the structures proposed by Ridgeway (Ridgeway, 2019). There may be many other games that are equivalent. The numbers describe payouts to (A,B).

reasoning (see quotes above), the primary argument in Ridgeway assumes that “modest rewards” are somehow given for deference, but these rewards are not theoretically well justified. Similarly to work in behavioural economics, intrinsic penalties or rewards can be added to the utility function to ensure equitable or fair behaviour (Rabin, 1993; Fehr and Schmidt, 1999; Akerlof and Kranton, 2000; Chentanez et al., 2005) (see Section 8 esp. footnote 58). I can explain this game theoretically as follows. I ask you, dear Reader, to imagine a modified prisoner’s dilemma game (see Section 5.1) in which a third option is added for each agent, that of “deference”. Ridgeway considers a solution in which one agent defers to the other leads to higher collective reward (this is reason for cooperation), but that lower-status participants in this solution receive proportionally less of the total reward gained by the group. I can write such a game in strategic form for agents A and B as shown in Figure 25(a). In this game, agents play a normal prisoner’s dilemma over their two actions cooperate and defect, get universally more if one defers, but do badly if both defer. There are three game theoretic equilibria. First, the both-defect solution is an equilibrium as usual and neither agent would be tempted to cooperate or defer if the other was going to defect (as he would do worse individually). However, two new pure equilibria are now present, in which one agent defers and the other cooperates. These two equilibria are *deficient* however, as the agents have no way of deciding which to choose. Ridgeway proposes to resolve this by adding a small reward to the defer action independently of the other agent’s action. This results in the game shown in Figure 25(b). Although the two pure equilibria are *strengthened* (in the sense that it is even more worthwhile to defer), the proposed game does not resolve the symmetry and the agents are no further ahead. One needs to assume further than these small deference rewards are allocated asymmetrically, based on the status evaluations of the actors (Ridgeway, 2019). Further details may be found in (Hoey and Schöder, 2020).

Instead, *BayesAct* proposes a resolution to this impasse with an identity-based attention mechanism. In *BayesAct*, and as pointed out by Ridgeway, a set of beliefs about relative status are held, and these break the symmetry according to the culturally defined labels assigned as identities to the participants. Although these identities may not be defined before an interaction, the first line of behaviour for humans is normally to establish them. They may be established through stereotypes about fixed status-valued attributes (e.g. based on gender or race), or on-the-fly signals of dominance or submission from a particular agent, in the “*language of status signs that others are likely to understand.*”(Ridgeway, 2019, p.56), based on the “*shared cultural stereotypes of the emotions associated with status*”(Ridgeway, 2019, p.57).

7.10 Lawler

Many social psychological theories about group formation and action can be discussed along one of two main lines (Lawler et al., 2009). *Relational* theories focus on the affective dimension of roles and identities as being the key factor in determining how macro-level social orders arise from micro (individual) level behaviours. People do things together because they feel like it and the interactions make them feel good. These interpersonal relations are learned and built over time, and they encode a social order. Individual self-interested actions tend to create disorder but can lead to new relational structures forming. Relations are strong and long-lasting because they are affective. Social identity theory (Tajfel and Turner, 1986; Hogg, 2006) is relational: it says that people select roles (identities) to play out and there are certain rules for each role that, if followed by all members, leads to a stable group structure.

Transactional theories, on the other hand, focus on how rationalistic humans following individual utility-maximizing strategies can coordinate to form coordinated groups. Transactional relationships tend to be short-lived and brittle, because they are replaceable: people don't care if one is replaced by another one. Social exchange theory (Emerson, 1981) is transactional. It claims that the driving force of group formation is individual goals and utility functions which are grounded in network structures. That is, situations are socially constructed to create individual incentives that rationalistic humans follow to ensure group coordination.

The theory of social commitments (Lawler et al., 2009) combines relational and transactional components. Lawler argues that modern society, becoming more and more influenced by economics and individualism, is moving towards an ecology of transactional ties, creating shallow, brittle and more dangerous (for the species) social structures (Lawler et al., 2009). His analysis carefully explores the space between these two views, and gives guidelines and arguments for how group processes can be built to take advantage of both. The key idea is that instrumental (transactional) commitments are the first line of defense in a new relationship, but that these commitments, if generative of positive emotions in a setting where responsibility is shared and contributions are non-separable, become "encoded" as relational (emotional) commitments, which are stronger and longer lasting. *Normative* commitments occur when transactional or relational commitments are externalized as norms and potentially encoded in the world somehow. A non-social illustrative example of such norms are "desire paths" in parks which start as instrumental actions by humans getting from point A to point B by cutting across the grass, and then are converted to normative actions as the desire path becomes encoded in the ground itself (as a path worn by many people doing the same thing) (Constant et al., 2018). A more social example are the rules governing order in meetings (e.g. Robert's), which are, in fact, reifications of the "correct" way of carrying out a meeting.

Research on communities of practice (Wenger, 1998) offers many situated examples of the theory of social commitments. The introduction of a new form into an insurance brokerage ("the B-F and G thing" (Wenger, 1998)) starts as a transactional relationship, but as the work group starts to use it, discuss it, reify it, the form takes on a relational meaning and becomes an integral part of the worker's practice. The transactional/relational divide is also seen in work on economic sociology, pointing to vertical integration in a firm which is normatively enforced and undersocialized, and to oversocialized theorizing in sociology which ignores economic actions of agents, and proposing an embedded approach that takes into account network structure (Granovetter, 1985).

BayesAct also combines relational and transactional ideas, by arguing that the core driving force behind human behaviour is relational and based on shared affective sentiments about identities and behaviours. However, cognitive *transactional* processes come into play when relations break down. Transactional relationships are set up to handle the breakdowns, but are rapidly integrated into relational processes if they are required to maintain the social order. I can therefore draw a parallel between instrumental ties and Durkheim’s *mechanical solidarity*, which I have done in Figure 2. Further, relational ties can be summarized as the Spinalien’s *organic solidarity*, also shown in Figure 2. Note that in Figure 21 the dimensions of mechanical and organic solidarity are pointing out from the origin, whereas here I am showing their projection onto the simplex created by increasing the missing variable.

7.11 Social Cohesion

Much attention has been paid to social cohesion, loosely defined as what ties a member to a group (Friedkin, 2004). While some researchers attribute cohesion to individual relationships, others point to network structures. Lawler’s approach (reviewed above) is to combine these effects, showing that one takes precedence over the other based on the amount of relationality in the group, coupled with the generation of positive emotions and the reification of the group. *BayesAct* combines these two elements as well, showing that both are a reflection of uncertainty management. As a group stabilises in membership and behaviours, it attains a certain level of relational commitments at a connotative level that lead to the reification of the group at the denotative level. This reification takes the form of the assignment of a label to the group.⁵⁵ This label then obtains a connotative meaning if has a positive association that is meaningful to each member.

Studies of social cohesion have also focussed on the impact of natural disasters on this construct. Recent studies of social cohesion in response to natural disasters suggest that cohesion emerges from within affected communities (Calo-Blanco et al., 2017), and is stronger for communities affected by natural disasters (Toya and Skidmore, 2014). Such reactions cannot be overcome by more information about the problem, but require increased certainty about the connotative status of affected people. Friedkin (2004) discusses social cohesion as arising from network structures, contrasting to identity-based social cohesion measures that are more individual-focussed. (Toya and Skidmore, 2014) discuss societal trust, one important component of social cohesion, in the face of natural disasters. They find some natural disasters in some countries increase societal trust, and this effect is somewhat generalizable, but depends on many factors that disrupt the correlation. For example, the size of a country matters, since natural disasters in large, low-density countries like Canada affect a relatively smaller proportion of people.

Similar ideas, in fact, can be traced back to Comte, who, in reaction to the social chaos resulting from the French revolution, attempted to bring a scientific dimension back into play in order to help move towards a more orderly situation (Comte, 1835).

⁵⁵ A personal anecdote demonstrating this was a group of high school buddies I once met while travelling, who, getting together every year for a week-long alumni trip, referred to themselves lovingly as “the knuckleheads”. This word reified their group, encapsulating a wide range of properties of the group, including memories of high school and of previous “knucklehead” trips.

7.12 Schwartz

Schwartz (1994) proposed a ten factor model of *values*, defined as five characteristics that define a value as a belief (1) pertaining to desirable states (2), that transcends situations (3), guides behaviour (4), and is ordered (5). These five characteristics can be reduced to two factors if considering a probabilistic and decision theoretic model: 1) active inference; and 2) transcendence of specific situations. Active inference computes a posterior over policies of actions (and thereby over futures), motivating action by attention as I have explained (1,2 and 4 and 5). To achieve transcendence, however, is more difficult as it requires (in a decision theoretic, denotative model) to capture *classes* of things, or Platonic *form* (characteristic 3). This then requires an intractable number of combinations of possibilities, as Schwartz says “*In order to cope with the reality in a social context, groups and individuals cognitively transform the necessities inherent in human existence and express them in the language of specific values about which they can then communicate.*” (Schwartz, 1994, p.21). The size of this “language of specific values” will be intractably large as it must contain a value mapping label for each possible context, including past and future.

With a connotative model, however, the problems are reduced, and these *classes* of things are simply *not* defined in the same subspace, but rather in a *parameter space*. That is, the values described in Schwartz (1994) are denotative labels associated with specific parameter settings in *BayesAct*. I can identify the three axis simplex I have been describing with the circumplex of Schwartz (1994), although the precise location of axes is more difficult as these values have no precise definition beyond how people respond to questionnaires.⁵⁶ Nevertheless, I can associate Schwartz’ *openness to change* with Bales’ *B* axis, and thereby to my $\hat{\alpha}$ apex. Similarly, *self-enhancement* I associate with Bales’ *U* dimension and thus to my $\hat{\delta}$ apex, Finally, *tradition and benevolence* are rooted in Bales’ *P* axis, my $\hat{\gamma}$ apex. These associations are shown in Figure 26 along with the Schwartz labelling of values in eleven different regions of the space (see Figure 26 caption for labels). Also shown in Figure 26 is the Bales space, from which correspondences to a wide variety of *personality* models (such as the five-factor model or the Myers Briggs indicators) can be made (Bales, 2020).

I will not analyze this correspondence in detail, but note a few key examples. In Schwartz’ regions I have numbered (7a) and (7b): social imposed expectations and normative closeness, are on the simplex at the $\hat{\gamma}$ apex ($\hat{\gamma}$ is high). Such groups are very homogeneous, a by-product of everyone following tradition. Individuals in these groups behave consistently with the norms of the group and therefore have little individual freedom. On the other hand, those seeking affectively pleasant arousal (3), are heavily focussed on connotative meanings, and seek pleasure, disregarding evidence from the world. Again, as in Figure 23, we see a relationship between more conservative political beliefs (in security and order and conformity, regions (8a) and (8b), and liberal ones (in affective pleasant arousal and novelty or mastery) in which individuality is highly prized. This corresponds to the relationships explored in Section 3.

Schwartz (1994) connects his circumplex of values to work on two-factor models by Eysenck, Fromm, Inglehart, Rokeach and Braithwaite, which I analyse here in the light of the simplex model.

⁵⁶Schwartz (1994, p.25) describes these as “fuzzy sets”, but this is not necessarily fuzziness in the sense of possibility theory.

Figure 26: (a) The Schwartz circumplex is mapped to the simplex in $\hat{\alpha}, \hat{\gamma}, \hat{\delta}$, with numberings corresponding to the regions as in (b) the original (Schwartz, 1994, p.24) drawing, with the Bales axes superimposed. The regions identified by Schwartz, numbered in both (a) and (b), are as follows (with the ten values, spanned in sets of two by the regions, in parenthesis following each one): 1) social superiority/esteem (power/achievement); 2) self-centered satisfaction (achievement/hedonism); 3) affectively pleasant arousal (hedonism/stimulation); 4) novelty and mastery (stimulation/self-direction); 5) diversity (self-direction/universalism); 6) enhancement of others (universalism/benevolence); (7a) social imposed expectations (benevolence/tradition); (7b) normative closeness (benevolence/conformity); (8a) preserve existing social arrangements (tradition/security); (8b) protection of order (conformity/security); (9) control (security/power).

- Eysenck (1954, Eysenck's) liberal/conservatism dimension is associated with Schwartz's openness to change (Schwartz 3 and 4) and conservation (Schwartz 8), which is exactly my A dimension. Schwartz also relates his region (1) power/achievement with Eysenck's "toughness" dimension and his region (6) with Eysenck's "tenderness" dimension. This corresponds to my D dimension.
- Fromm (1949) discusses humanistic vs authoritarian, which map to Schwartz's universalism and self-direction (5) to security, power and conformity (regions 7b, 8, and 9). This also corresponds roughly to the A axis in the simplex model.
- Inglehart (1977) discusses materialist vs. post-materialist, associated with Schwartz's region 9 to his region 5, corresponding to my $\widehat{DA}$ axis.
- Somewhat more difficulty was found by Schwartz in relating to Rokeach (1973). However, these difficulties, as well as those he had with Braithwaite (see next bullet point), are primarily caused by these authors leaving out one of the three dimensions that define the simplex (on which Schwartz was working). For Rokeach, equality and freedom are two orthogonal dimensions, but Schwartz is puzzled that these lie next to one another on the circumplex (as universalism and self-direction, respectively). However, once equality is associated and freedom are noted to be perpendicular to the simplex, then they are somewhat independent of any variation thereupon. This places self-direction and universalism (corresponding to freedom and equality) near one another on the simplex, somewhere in the middle depending on the precise weighting. My guess is that the confusion surrounding the missing dimension stemmed from Rokeach using dimensions of republican equality and negative freedom, which Schwartz misinterprets as negative equality and negative freedom, which are contradictory.
- Similar issues were found with Braithwaite (1994), and these can be resolved by associating Braithwaite's *international harmony and equality* (IHE) with my $\hat{\gamma}$ axis, while Braithwaite's *national strength and order* (NSO) with my $\hat{\delta}$ axis. Braithwaite essentially is missing the $\hat{\alpha}$ axis, and thus Schwartz's associations between IHE and his universalism, as well as between NSO and security, are not quite correct. Rather, NSO, considered as $\hat{\delta}$, maps to Schwartz's region 1, while IHE, considered as $\hat{\gamma}$, would map to region 7a, and corresponds to the $\widehat{\Gamma D}$ edge in the simplex. Schwartz's associations means he is working roughly along the A axis, but trying to project it along the $\widehat{\Gamma D}$ axis. As these are nearly perpendicular, the projection gives rise to only weak negative correlations, as noted by Schwartz (Schwartz, 1994).

Charles Tilly also works from a two-dimensional space (Tilly, 1998, p.54), in which *scripting* corresponds to $\hat{\delta}$, while *shared local knowledge* corresponds to $\hat{\gamma}$ and $\hat{\alpha}$. Scripts are instrumental definitions of situations, and are used to launch new social interactions (Tilly, 1998, p.57). Shared local knowledge then is solidified and relational interactions start to occur, leading to what Tilly calls *deep improvisation*, which is a highly trusting group with a lot of shared knowledge (so high in both $\hat{\alpha}$ and $\hat{\gamma}$ but low in $\hat{\delta}$). Tilly's *intense ritual* with lots of scripting and shared local knowledge will be difficult to maintain as a social system, while *shallow improvisation* (little scripting or shared local knowledge, as "pedestrians on a crowded sidewalk" (Tilly, 1998, p.54)) will be brittle and non-functional as well. The other two edge cases are then *thin ritual*

(lots of scripting, little shared knowledge) and *deep improvisation* are both workable solutions with different outcomes.

7.13 Self-Affirmation Theory

Self-Affirmation Theory (Sherman and Cohen, 2006) starts from the premiss that a person’s “self-integrity” is a major force behind people’s responses to situations, particularly threats. The idea is very much in line with the *BayesAct* view: maintenance of the self is a primary concern for people, and when the self is affirmed, they focus on the informational value of the situation instead of on the integrity of self. Self integrity is defined somewhat broadly, and includes, roles, values, group identities, beliefs, goals and relationships (Sherman and Cohen, 2006), all of which are accounted for in *BayesAct* (see the start of Section 7).

Sherman and Cohen (2006) review many studies that demonstrate the effects of self-affirmation, including cognitive dissonance (Festinger, 1962) (for a *BayesAct* interpretation of this specific case, see (Hoey et al., 2020)), and political polarization (see (Facciani, 2020) for an affect control theoretic interpretation as well). In particular, people exposed to more connotative uncertainty about the self will tend to focus more on prior denotative meanings, and will tend to be less open minded (stuck on identity-based stereotypes) and less able to learn (explore) the denotative structure of the environment. A series of studies by Lord et al. (1979) showed that people evaluated scientific results more objectively if they were in line with their prior beliefs. The participants were more likely to find flaws in a study if it contradicted their prior beliefs. From a self-affirmation standpoint, this happens because their self is threatened, leading them to focus more on anything that will affirm the self (here it’s denying evidence). An extension of this study primed participants with self-affirmation or self-uncertainty (Cohen et al., 2000). Those in the affirmed condition were even more likely to objectively evaluate scientific studies, even if it countered their beliefs, while those whose self was primed to be more uncertain would be less likely to do so. Those who were affirmed were also more likely to shift their beliefs towards the evidence. Those non-affirmed participants then fell back to denying evidence that did not support their selves, as in (Lord et al., 1979).

In *BayesAct* terms, the posterior connotative distributions for the self-affirmed group are assumed to be less dispersed, leading to a decreased focus on the denotative posterior. Those who were not affirmed have more dispersed connotative distributions, leading to increased focus on denotative observations that confirm connotative meanings about the self. Let me go over that a second time. The posterior connotative distribution is the distribution over EPA space of the connotative meanings of the situation, including an identity for the self, for the experimenter, and for the expected action. If a participant is *not* self affirmed, they are likely to be seeking some fairly neutral identity like “research participant”, or “student”. If they then read an article about a moral value and are asked to rate its validity, they would seek some dimension on which to interpret this situation, and would likely select as most salient an identity describing their political polarization (i.e. “I am a liberal”). However, this would be combined with the weaker identity of “research participant”, and serve to further distribute the connotative posterior. In such a case, the participant will be more focused on the denotative meanings of being “a liberal”, for example, and would find it more difficult to accept evidence that was not in line with this. In the self-affirmed case, the connotative posterior is precise, leading to increased dispersion in the denotative state through the somatic transform. In this case, the search space is broadened,

and a wider variety of solutions can potentially be found. One would expect that the shift in belief toward the evidence having validity would percolate up to the connotative state eventually, potentially even shifting this state significantly. The most interesting thing about the *BayesAct* view is that it brings in affective meanings as a key driving force, and gives a more parsimonious interpretation of these results as emotional reactions of self-defense in the presence of denotative uncertainty.

Interestingly, Paxton et al. (2012) show something quite similar, in which people are presented with a morally challenging situation (e.g. involving incest), and their inability to reason correctly about bad evidence presented does not stop them from making a choice about the situation. However, when given time to slow down (2 minutes pause), the participants lost this inability, and were more accepting of the morally difficult but logically correct situation. In self-affirmation terms, they were placed on high connotative alert with the story, and thus were able to discount evidence that did not favor it. As this faded over the two minutes, their objectivity returned and they were more open to the logical reasoning.

7.14 Personality Theories

There are a number of theories of personality in psychology that maintain the notion that an individual has some persistent, long-term, internal character traits that define a particular way of acting in the social world. These characteristics are properties of the individual, and define him or her, and are carried from one situation to the next. For example, the five-factor model describes personality traits in a five dimensional space, with features corresponding to extroversion, agreeableness, neuroticism, openness, and conscientiousness (McCrae and John, 1992). The other major model is the Myers-Briggs (Myers and Myers, 2010), as exemplified in the so-called “16 personalities” score (see 16personalities.com). These models are known to be related, although the five factor model seems to be the more validated of the two (McCrae and Costa Jr., 1989).

The five-factor model characterises personality in a five dimensional space, where each dimension is usually discretized at some level. The dimensions are openness, conscientiousness, extraversion, agreeableness and neuroticism. Bales has mapped these five dimensions to his space, noting that two of the personality dimensions are opposing ends of his axes (Bales, 1999).⁵⁷ Bales connects neuroticism, for example, to his NB direction. This corresponds in *BayesAct* to high $\hat{\alpha}$, low $\hat{\gamma}$ and $\hat{\delta}$, roughly to his region 10 (anti-authority opposition, see Section 7.3). When $\hat{\gamma}$ is low, the somatic transform is weak, and in more ambiguous situations (where the denotative dispersion is high) the connotative bias will shift the posterior towards over-estimates of risk (i.e. distributions that are less informative). It is known that neuroticism corresponds to perceptions of increased risk in ambiguous environments (Satchell et al., 2020), thus confirming the relationship. Similarly, *extraversion* is mapped to Bales’ U dimension, corresponding to high $\hat{\delta}$, but low $\hat{\alpha}$ and $\hat{\gamma}$. *BayesAct* agents with these parameters will be information seekers, focussed on denotative meanings (are operating with wide beams searching for small targets), and thus will be driven to interact more with the world, and will be perceived as more extraverted. Similar associations can be made between *conscientiousness* and α (low $\hat{\alpha}$), between openness and low $\hat{\delta}$ (mid-range $\hat{\alpha}$ and $\hat{\gamma}$, roughly my *D* axis), and between agreeableness and $\hat{\gamma}$.

⁵⁷I have not yet found a paper in which the mapping is explicitly spelled out, but this webpage <https://www.symlog.com/SYMLLOG/SYMLLOG-NeoPi.aspx> shows it (accessed August 2020).

Some social theorists, however, tend to reject the notion of a stable personality (the “myth of individual persons” (White, 1992, p.207)), arguing that a person is defined by the social situation they are in. As White puts it, “*personality psychology remains grotesqueley naïve about the constituents of ‘everyday life’*” (White, 1992, p.206). According to Burt (1992), individualist personality ignores social and network constraints, making it a very limited view of human behaviour. Berger and Luckmann (1967, p.174) refer to them as “collective identities” in a similarly disparaging manner. According to these sociological views, markets make value networks of firms and actors, and “structural holes” (unconstrained links) in the network lead to entrepreneurial opportunities. Constraints in the network, on the other hand, can be handled using one of three strategic actions: to *withdraw* from the network, to *expand* the network, and finally to *embed* the network in some other network (Burt, 1992, p.264). If embedding is the strategy of choice, then new relationships must be defined, and this leads to the “emotional residue” that we label as either “personality” on an individual level, or “firm” on a organization level.

“[Personality] is the emotional residue of managing concrete relationships with specific individuals.” (Burt, 1992, p.264)

“...social and emotional organization as a kind of residue that accumulates in the wake of entrepreneurial players navigating around the constrained relations that define a market” (Burt, 1992, p.229)

“...today’s institutions are the residue of yesterday’s entrepreneurial efforts” (Burt, 1992, p.229)

A *BayesAct* formulation agrees with the incorporation of “everyday life”, as it holds that different institutions may lie in different regions of the simplex of Figure 2. Based on the context, or what institution is currently “in play,” a particular uncertainty tradeoff is selected, and a particular personality is apparently displayed. It is, of course, easier for an agent to maintain fewer of these models, e.g. by using the same uncertainty tradeoff (model parameters $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$) for different institutions. However, agents whose contexts are shifted (e.g. they move to a different culture, or start hanging out with a new set of friends), will adapt their models to fit the new environment, shifting their personality in a way that encourages exploration (that “gets” action) but that results in a return to equilibrium (that eventually “blocks” action). A similar argument is made in attitude representation theory, which argues that an individual will respond similarly in situations that have a similar “matching” between objective and subjective representations (Lord and Lepper, 1999).

In *BayesAct* therefore, an agent uses some particular model tradeoff $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$, and can move, perhaps only slowly over larger distances, from one to another based on the context. For personality models, the person cannot shift, and remains static at some particular location. Not only does this type of static theoretical approach preclude people from changing cultural niches (thereby enhancing diversity), it causes major problems for all group processes, because it tends to stymie natural small shifts in parameters that may be obvious and easy to do, and that may induce much more efficient group processes. In fact, it may have impede *cultural adaptability*, a facet of a person that is increasingly recognized as important for organizational efficiency (Corritore et al., 2020). Personality models therefore reinforce a static consistency of people across time, thereby requiring a more static view of the environment, something that, as

I have shown, is more closely related to conservative political views than to liberal ones.

7.15 Morality

Moral psychology has made major advances in the study of how people make decisions and how social orders are thus defined (Haidt, 2012, 2001). Embracing a two-system view, moral intuitionism holds that a fast thinking “system 1” makes a quick decision of what to do based on a moral code defined by the society *and the context*, and this is potentially refined (given sufficient resources and time) by a “system 2” rational thinker. The quick thinker is usually held to be a large and powerful motivational force, the course of which the rational thinker can alter, but only with difficulty (Haidt’s “elephant” and “rider,” see (Haidt, 2012)).

Shweder (1990) categorises moral discourse into three “codes”, given the names “autonomy”, “community” and “divinity”. These correspond precisely to my three axes of validity, coherence and dependence, respectively. It may be argued that these moral concepts line up with the three axes of solidarity instead (such that autonomy=mechanical, community=organic, and divinity=artistic), however in either case somewhat the same analysis can be carried out (albeit results in two analyses, they will be isomorphic, and therefore redundant). The moral ethic of autonomy sees the individual as having free choice, or perfect negative freedom, but giving up positive freedom in doing so, in that each individual’s actions are defined (e.g. as rational). Thus, autonomy is my “validity” axis ($\hat{\delta}$), or alternatively, my A axis. That of community sees the group as more important, and so defines roles and role structures for people to live with. They discount free choice, but open up opportunity. Thus, community is my “coherence” axis ($\hat{\alpha}$), or alternatively my G axis. Finally, the moral ethic of divinity focusses on the *connection* between the individual and the group, and holds that all must hold true to some spirit or essence. The ethic of divinity connects actions to individual strongly and followers think its important to follow what is important in each moment. They end up with a very deontological ethic, which forces them to all be the same under the same divinity (for how can there be two?). They are therefore high on the $\hat{\gamma}$ axis (or high on the D axis).

Therefore, we find societies that are more liberal showing more community and autonomy ethic (more organic solidarity), while those more conservative showing more autonomy and divinity ethics (more mechanical solidarity). The key is that the two sides simply cannot get along. From a free energy perspective, group C makes rules that increase the overall energy by pushing group L “off the simplex.” Similarly, group L makes rules that push C off the simplex. That is, any rules made by one side cannot be stomached by the other. Therefore, a democratic system in which national choices are made by a partisan government is non workable and is bound to fail. This finally leaves the question of whether there is a democratic system that is workable and I have already alluded to it in Section 6.3. I will discuss this further in Section 9.7.

8 Artificial Intelligence

Outside of cultural sociology and social psychology, the quest for a denotative agent continues, and has resulted in large scale mathematical models, implemented on large scale computers, and has now co-opted the term “artificial intelligence” (AI) as a somewhat narrow definition of an agent that performs well in a particular game, e.g., a video game (Silver et al., 2017). However,

there is a known lack in AI of an ability to process connotative or emotional meanings, something taken up by the field of “affective computing” (Picard, 1997). Without these connotative meanings, artificial intelligence defined denotatively may be “intelligent” in some narrow domain, but not so in the social way that even animals are (Darwiche, 2018; Haslam et al., 2008). The original goal of AI was to understand intelligence by trying to build it. This remains my goal, and as the term “artificial intelligence” has been co-opted by commercial interests to mean anything that solves some specific problem better than a human, I prefer to use “computational human Intelligence,”, which can be shortened to “computational intelligence.”

In a critique of AI, Adnan Darwiche claims that even the AI researchers that try to stay true to the original goals of AI, are still basing many of their theorems on incorrect axioms, or that many of their deductions are based on incorrect premisses (Darwiche, 2018). This is exemplified by Darwiche’s statement that “Every behavior, intelligent or not, can be captured by a function that maps inputs (environmental sensing) to outputs (thoughts or actions).” This is simply not true, because the function must also map outputs to inputs. Managing a function that does both simultaneously is substantially more difficult than only handling one direction, but is a requirement for intelligent agency (Friston, 2010).

My own work in computational intelligence has been primarily focussed on unifying emotional and cognitive reasoning. By proposing the *BayesAct* model, my co-authors and I have sought to unify the comprehensive and mathematical affect control theory with traditional decision-theoretic models of reasoning under uncertainty and proposed it as a class of computational intelligence models (Hoey et al., 2016; Asghar and Hoey, 2015). This class of Bayesian models have shown relevance to social-psychological theorizing (Schröder et al., 2016; Hoey and Schröder, 2015) and to empirical results in sociology (Freeland and Hoey, 2018).

The most relevant recent work in artificial intelligence is focussed on multi-agent reinforcement learning (MARL), a problem in which multiple agents attempt to solve a common problem. There are many flavours, such as agents with a central controller, agents who can communicate, agents who share utility functions. Reinforcement learning methods are normally decision theoretic, and therefore denotative ($\max \hat{\delta}$), and are thus unable to handle social situations. To do so, they require some emotional reasoning.

8.1 Reinforcement Learning

Connections of emotions with reinforcement learning have been explored by a number of authors (Moerland et al., 2017; Hogewoning et al., 2007; Marinier III and Laird, 2008; El-Nasr et al., 2000; Broekens et al., 2015). In traditional reinforcement learning (RL), an agent is tasked with both learning about his world (primarily the utility of situations), and of acting in the same world. The basic quandary for the agent is whether he should *exploit* his current knowledge of the world, or *explore* something new in the hopes of discovering even better situations: a difficult tradeoff indeed between a sure bet and a random chance. Many RL agents tackle this tradeoff using some form of *optimism under uncertainty*: if something has not been tried (or has not been tried for some time, or insufficiently many times), assume it will lead to high utility outcomes (Brafman and Tennenholtz, 2002; Kocsis and Szepesvári, 2006). This method works in practice because it ensures sufficient exploration, and agents don’t get “stuck.” In traditional RL, exploitation is seen as a cognitive/rational skill requiring (usually intense) computation, since it involves predicting the future based on learned knowledge, and analyz-

ing the costs and benefits of different strategies. Exploration, on the other hand, is seen as something that could be guided by any number of (possibly affective) elements. For example, Hogewoning et al. (2007) equate higher valence (which is equivalent to reward being higher than expected, so things are going well) with increased exploitation of current knowledge, whereas lower valence/reward (things are going worse than expected) pushes an agent to explore. This view is largely consistent with the *affect-as-cognitive-feedback* view (Huntsinger et al., 2014) where higher valence facilitates usage of existing mental constructs, whereas negative valence inhibits (and thus forces an agent to seek new solutions through exploration). In general, expected and immediate emotions are hypothesized to be related to expected utility and modify action choices accordingly (Loewenstein and Lerner, 2003). In particular, emotions such as fear/hope/joy/distress have been related and computed directly from value functions and rewards. For example, “joy” is defined as the likelihood of a change in expected value, while “hope” is defined as expected value (if positive) (Broekens et al., 2015). Gomez and Insua use a decision theoretic definition of hope, fear, happiness and sadness, in much the same way (Gomez Esteban and Rios Insua, 2017). These emotions are computed based on expected utility, and the elicited emotions’ intensity is partially based on uncertainty. If a threshold in uncertainty is crossed, then a short-circuit “impulsive behaviour” is triggered that overrides the rational one. Although the authors claim this is a “System 1/System 2” dual process approach (Kahneman, 2011), it is much more in line with the Simonesque “interrupts” approach (Simon, 1967).⁵⁸ A recent survey of the connections between emotional appraisals and elements in reinforcement learning is in (Moerland et al., 2017).

⁵⁸A generalisation of these approaches is intrinsically motivated reinforcement learning (IMRL), which is an evolutionarily plausible mechanism for allowing agents to learn a reward function that will lead them to be maximally fit (Chentanez et al., 2005). The idea is to search the space of reward functions for one that maximizes the agent’s fitness as defined by the extrinsic (usual, designer-provided) reward. A distinction is made with normal “reward shaping” because of resource bounds or agent limitations: the optimized reward function in IMRL will take these bounds into account, finding the best possible reward function given the limitations of the agent. Emotions have been proposed as defining the space of reward functions over which an IMRL method will search (Sequeira et al., 2014). A set of emotional appraisal features are defined, the weights of which are exhaustively examined to determine the optimal setting for a given domain. The appraisal features used in (Sequeira et al., 2014) are novelty (defined in terms of how many times a state-action pair has been tried), goal relevance (defined as a heuristic estimate of the distance to a maximally valued state), controllability (with uncontrollability measured by the Bellman error), and valence (measured by the value function directly). These appraisal features resolve to standard heuristic methods for guiding reinforcement learning, such as *optimism under uncertainty* (exploration bonus) or Bellman error (exploitation bonus), being now referred to as an emotional appraisal of *novelty* or *control*, respectively. Even in the multi-agent case (Sequeira et al., 2011), the appraisals are direct encodings of altruistic reward functions, leading to simple modifications to reward functions that cause rational agents to be more cooperative (as in (Nowak, 2006)). The learned reward functions are domain-specific, and fail to generalise. Attempts to find an optimal “universal” reward function result in an agent that optimizes the extrinsic reward, with a small exploration bonus (encoded as a negative reward for controllability) that decreases as the agent learns the environment (Sequeira et al., 2014). In (Marinier III and Laird, 2008), the SOAR cognitive architecture is augmented with a reinforcement learning agent that uses a synthesis of many emotional appraisals (termed a “feeling”) as a reward signal. As in the above approaches, these emotional appraisals are direct encodings of decision theoretic primitives or heuristics (e.g. direction to goal). Although Marinier and Laird (Marinier III and Laird, 2008) associate this approach with IMRL, it is not related as the agent does not attempt to match the external reward function but only uses its internal reward signal to learn from.

8.2 Relating Uncertainty and Emotion

Related views of emotion include the identification of negative valence with increased uncertainty (FeldmanHall and Shenhav, 2019) or with change in free energy (Joffily and Coricelli, 2013), or expected free energy (Hesp, 2018). One component of identity (esteem) is used to modulate a reward function in (Moutoussis et al., 2014b) in order to make cooperation the more salient policy in a social dilemma. However, in (Moutoussis et al., 2014a), interpersonal relationships in general are linked to heuristics that facilitate approximate Bayesian inference, allowing agents to circumvent the intractability of optimal Bayesian inference. Importantly, the models of self and other exist specifically to influence agent actions.

In a similar vein, (Jaques et al., 2019) consider social *influence* as an additional factor in an agent’s utility function in a sequential social dilemma, and the relative weights of this factor can be learned by an agent about other agents. Influence is a signal of power, and we can therefore see the same concept being applied in this multi-agent reinforcement learning context. Note that in earlier work (Ray et al., 2008), the additional factor in the utility function is parameterized by *envy* and *guilt*, which map to *both* power and evaluation dimensions.⁵⁹ Importantly, the precision of the estimate of the other’s type is critically important (Moutoussis et al., 2014a), especially as ambiguity grows (FeldmanHall and Shenhav, 2019). In a trust game, cooperative solutions are linked to valence (self-esteem) in (Moutoussis et al., 2014b), much like the tendency towards fair solutions in uncertain situations in (van den Bos, 2001; Hoey et al., 2020).

Relating uncertainty or model precision with affect opens up a new type of dual-process model that has recently been explored. Hesp *et al.* (Hesp et al., 2019) describe a deep active inference model in which “affective charge” is a measure of an agent’s inability to predict the future, and feeds directly into the precision of policy search. Thus, agents with models that are not working well (are not bringing the agent to expected states in the future) naturally expand search to more diverse policies, while agents with models that are working well increase policy precision and so stick with what is working now. However, affective charge does not take social factors into account, and so relates more strongly to the lower level of core affect. Nevertheless, the integration of *BayesAct* into such an active inference model is the subject of ongoing efforts. In related work, FeldmanHall and Shenhav (FeldmanHall and Shenhav, 2019) describe an abstract Bayesian model in which humans are motivated to reduce uncertainty in their distributions over their own actions, conditioned on their current appraisal of their state. The ability to reduce uncertainty has clear evolutionary advantages, meaning humans have an intrinsic motivation to reduce it. Affect acts as an easily accessible measure of uncertainty over actions for the agent, and provides an error signal that can be used in a feedback loop to reduce uncertainty. As uncertainty grows, e.g. in a social situation, negative affect is created and an agent is motivated to reduce this, restoring positive affect, by taking actions that reduce uncertainty. In social situations, the state of the world includes the states of other agents, such as their traits, goals and emotions. A continuum of strategies, from “automatic” control based on stereotypes and impressions, to “controlled” processes based on perspective-taking and effortful search, are used to reduce the uncertainty. This continuum maps to a Bayesian tradeoff between prior and evidence, and is well described in the *BayesAct* model by a dual-process system I explained in

⁵⁹Envy is a combined emotion of sadness and anger, which therefore are mixed between negativity and dominance (aggression), while Guilt is a combination of joy and fear, mixing positivity with submission (Plutchik, 1980).

Section 4.

These approaches show how valence (and arousal in (Hesp, 2018) and power in (Jaques et al., 2019; Ray et al., 2008)) may be related to uncertainty, more precisely to the precision of policies. Affective responses are proposed as resulting from active inference in the conceptual model of (Smith et al., 2019). *BayesAct* shows how this relationship can be linked to social psychological theory, providing a bridge to sociological analytics, by associating dimensions of sentiment with affective response representation in sentiments of *happy* vs *sad* (Evaluation) and of *angry* vs. *afraid* (Potency). However, the sentiment dynamics of *BayesAct* exist at the affective response representation and conscious access levels in the three-factor active inference model of (Smith et al., 2019), and would be involved in conscious and subconscious interpretations in working memory (Smith et al., 2018). In fact, in the precursor model of Lane (Lane, 2000), the level 5 theory is “yet to be formulated” (p362), and “would involve social cognition and would focus on how differentiated awareness of self and other influences social behaviour...” (p362). More recent work has noted that “the theory of mind abilities associated with the fifth level of emotional awareness” are not yet fully addressed (Smith et al., 2019). I suggest *BayesAct* exists at this level, which can handle “blends of blends of emotions.” For example, a feeling of happiness and sadness generated by thinking about a lucky friend feeling happy and proud, but a bit worried about his effect on me. This was further explored in Section 5.3.

BayesAct agents will naturally shift in response to the unpredictability of their environment, but may do so starting from different individual “set points,” defined by the parametrization of the *BayesAct* model. These differences in the management of uncertainty have been noted before in work demonstrating individual differences in a bias-variance tradeoff at the perceptual level (Glaze et al., 2018). In keeping with a hierarchical Bayesian model of the mind, I have shown how this tradeoff can exist at the level of conscious cognitive constructs.

8.3 Ethical Considerations

When building intelligent agents, and especially those with socio-emotional capabilities, ethical issues must also be carefully considered. The moral machine experiment (Awad et al., 2018) showed that people have shared behaviours as moral decision makers, with consistency across, and diversity within, a culture. We present a possible model for this in *BayesAct*, with this consistency arising in a sentiment (connotative) space with a simple prior distribution. This connotative space and associated temporal dynamics has a direct multimodal (emotional) communication channel providing it with information, and is learned through interaction with a social group. With a connotative space and dynamics which are consistent with others’, agents can benefit by having easier focus on some aspects of the denotative world, specifically those that are relevant as solutions to social dilemmas. When they are able to follow these prescriptions for dilemmas, they become “members” of the social group in which they are learning.

Thus, any moral decisions made by the agent would be consistent with those made in its social group, and therefore be more acceptable. Inconsistent agent behaviours result in the ostracism of offending agents, communicated with emotional signaling and less cooperative behaviour. In theory, therefore, this may be possible but as indicated above is fraught with ethical dilemmas that I do not make any attempt to resolve here.

Social engineering and Social Robot Engineering are emerging fields that combine robotics and AI with security analysis. Social engineering is already prevalent, and requires an emotional

connection to the person being engineered. These issues are discussed in depth by Postnikoff (2020), but authority, conformity, reciprocity, empathy and trust are key elements to the definition of the problem. With more well defined models of social agents, we are running into the dilemma posed by these security and privacy concerns.

9 Dyamics of Social Networks

The previous sections have discussed in some depth the relationships between the management of uncertainty in groups, defined in a three dimensional parameter space I have denoted $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$, and a number of sociological theories that categorize groups into two- or three- dimensional spaces. However, the most interesting aspect of these associations is they provide an analytical tool that can be used to look at how social networks change over time, under the actions of the agents in them. The modeling involved in doing so is significantly more complex as it involves many more agents than three, and also requires handling agents with different $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$ settings.

The difference between real world groups lies in how much uncertainty is handled connotatively vs denotatively, defining a scale between underfit models that rely on connotative representations and result in high bias, low variance predictions and behaviours (I called these “L” agents in Section 2.1), to overfit models that rely on denotative representations and result in low bias, high variance predictions and behaviours (“C” agents). What kind of model is used by a social group depends on the learning environment, which, in a self-referential fashion, depends on the types of agents that populate the group. For example, more diverse populations of “L”s will “force” member agents to use higher bias models, as the uncertainty in the denotative state is much higher. “L” agents using higher bias models will be more tolerant of diversity, allowing for heterogeneity in the group. “C” agents, on the other hand, will require more homogeneous groups, and rely more on deliberative and observational processing. Further evidence of this relationship can be found in the studies of Van den Bos (van den Bos, 2001) who showed that induced feelings of uncertainty led to increased positive affect caused by perceived fairness. As diversity leads to uncertainty, the higher bias agents will favor fair (culturally normative) solutions over less equitable ones. Higher variance agents, on the other hand, will be more tolerant of inequality. In Hoey et al. (2020) we further provide a model of this effect using *BayesAct*.

In this section, I sketch some ideas for how these dynamics can be handled, and I consider a simple example of a hiring process, as discussed in Burt (1992), and then discuss other aspects of ambiguity, polarization, and automatic government.

9.1 Changing structures

There are two ways in which the handling of uncertainty in a group can be modified (shifting parameters $(\hat{\alpha}, \hat{\gamma}, \hat{\delta})$), both of which may lead to changes in the social structure. First, environmental change (exogenous) can “force” the group in one direction or another denotatively. Second, agents wishing to join a group will be required to shift their usage of uncertainty biases in order to better fit with the group’s definitions and be “accepted” (Burt, 1992), see also Section 4.3. An agent’s handling of uncertainty is therefore context dependent, in that it will be different in each group. The differences are both because of changed identities (e.g. a pro-

fessor at work, a father at home), but also because of changed uncertainty management that arises from the particular social structures at play in this particular group, some of which may be determined by exogenous forces (e.g. physical barriers or catastrophic events). An agent’s admittance into a group will shift this balance, which will need to be re-negotiated by all agents. However, due to the statistics of interactions, the new individual will interact often with agents already “in” the group, while the existing group members will interact with the newcomer infrequently. Thus, the newcomer will change more than the rest of the group, although it may be possible to get the group to make significant changes as well. Social entrepreneurs are defined as people who modify the network structure (e.g. by making or breaking a link) in order to shift the group towards a “setting” that is more in line with the entrepreneur’s individual goals.

For example, consider getting a job as a canonical example of an “outsider” becoming an “insider” (joining a group). It is well known it is helpful to have a strong sponsor in a firm with ties to the boss doing the hiring (Burt, 1992). This triadic dynamic can be used to set up a relationship of trust between the job applicant and the boss. In order to hire the applicant, the boss needs to be sure the applicant, if hired, will finish the job. The boss can base this on fact, or on the sponsor’s “word”. Fact and hearsay have significant weight in this case, and hiring one applicant over another may come down to the sway of the sponsor. In either case, once hired, the employee and boss naturally enter into a trusting relationship, and so the uncertainty management of both employee and boss will shift to fit better with the new social structure.

Now consider what happens if control is threatened by the new employee. Connotative precisions are decreased because of a change in the status order (there is more uncertainty created in this status order), and so the posterior shifts towards the denotative explanation. However, due to the vast complexity of any real situation denotatively, a denotative solution to a group problem will often be handled through an authoritative “forcing” of structure and hierarchy. Of course, there are also many examples of novel groups that self-organize without imposition of formal hierarchy (such as groups of students, researchers, or open-source software developers with egalitarian values), but according to prior affect control theory simulation research, such egalitarianism causes more stress and affective deflection in task groups than hierarchical solutions (Hoey et al., 2018; Morgan et al., 2019).

Figure 27: (a) Boss-C-Sponsor triad is 120C (see Figure 13); (b) Boss-C-Sponsor triad is 210 after C is employed (see Figure 13).

I will now consider the case of the job applicant in more detail. There are three actors in this group, and they form a triadic relationship as shown in Figure 27(a). The actors are the hiring *boss*, the supportive *sponsor* and the applicant C (denoted as a variable because it will change).

Prior to the hiring process, the agent C is viewed as a *stranger*, C and the *sponsor* have affinity (support/like) each other, while C has affinity for the *boss*, and the *boss* has affinity for the *sponsor*. I will approach this problem in two ways, first denotatively, then connotatively. I will then discuss what a tradeoff between these two solutions would like like. I will focus on the interpretation of the situation by the *boss*, who has the primary action to take in the situation (to hire C or not). Thus, *boss* has two options for action: to $a = \text{hire}$ $\mathbf{f}_b = (\text{EPA} : \{2.2, 2.4, 0.0\})$, or $a = \text{reject}$ $\mathbf{f}_b = (\text{EPA} : \{-1.4, 1.3, 0.36\})$.⁶⁰ *boss* maintains a denotative frame representing C , with slots for a denotative label for C , $x_c = \text{stranger}$, $f_c = (\text{EPA} : \{-0.15, -0.18, -0.20\})$. *boss* maintains a similar frame for the self, with $x_a = \text{boss}$, $f_a = (\text{EPA} : \{0.89, 3.08, 1.33\})$. Finally, *boss* has a representation of the *sponsor*, probably with a positive and powerful, but inactive, label such as $x_c = \text{benefactor}$ $y_c = (\text{EPA} : \{2.1, 2.2, 0.1\})$. *sponsor* and agent C also have representations of the other two agents, as will agent C , but I will focus here on *boss*.

9.1.1 Denotative Solution

Denotatively, the *boss* will have some representation of the proposition that there is a job, j , that requires doing, and that there is a worker, w , who is assigned to that job. Further this worker has a competence, o_j , which I will take to be a real value in $[0, 1]$ denoting the probability of success should this worker attempt job j . The success or failure of the job, denoted by Boolean variable J , with $J = \text{true}$ indicating job success, is dependent on the competence of the worker. Suppose that the job's probability of success is 0.9 should the worker be competent, and 0.5 if not. Now the *boss* needs to consider applicants. Each applicant can be represented by a frame giving an estimate of the competence of the worker. Initially, suppose that all competences are set to 0.2, meaning that the *boss* has a low expectation from applicants. Thus, should the *boss* employ any worker sight unseen, his expectation is that the job will succeed with probability $0.9 \times 0.2 + 0.1 \times 0.8 = 0.26$. Thus, his expected reward is $0.26 \times r_j$, where r_j is the reward he gets for successful completion of job j .⁶¹

Now suppose that the *boss* receives some information about one particular applicant, C . The *boss* is in a position of deference with respect to the *sponsor*, from whom the information originates. The *boss* has previously acted upon advice from the *sponsor*, and has found that each time the advice is good. That is, of the set of applicants that the *sponsor* has recommended in the past, most have proven to be competent for the job they were hired for. Therefore, the set of applicants which have the “word” of the *sponsor* all are assigned a higher competence, leading to a higher probability of success, and then to a hiring decision being swayed in their favor.

Through the story of this positive hiring decision, agent C transitions from denotative identities of *stranger* to *applicant*, then to *competent applicant*, and finally to *employee*. The belief of the *boss* that C is competent shifts throughout this process, and is then confirmed (or disconfirmed) by the actual performance of C . Further, the *boss* may have some stereotypes that he assigns to the applicant based on observable status characteristics such as race or gender. These stereotypes may have a denotative influence on the estimates of the job getting done. For example the *boss* may have a denotative belief that men are more likely to work overtime hours, which may sway his estimates of competence and hence job completion.

⁶⁰ All EPA values in this section taken from the USA 2014 study.

⁶¹ There may also be costs for *not* completing the job, and these would change the dynamic significantly.

9.1.2 Connotative Solution

Connotatively, the identities start as *boss-stranger-benefactor*, and the two actions for the *boss* of hiring and rejecting have deflections of 1.0 for the behaviour *hire* and 1.8 for the behaviour *rebuff*,⁶² so he'd hire the stranger since rebuffing him is so negative. However, these connotative solutions would be combined with the denotative solution above, which would rate the stranger's probability of success as very low with great certainty, and therefore this would "override" the connotative differences above.

However, upon receiving word from the sponsor, the *boss* can either *listen to the sponsor* (deflection is low, 1.9), or can *ignore* him (deflection is high, 7.8). So the *boss* would definitely *listen to the sponsor*. What does this entail as far as the final action goes? If there is no information from the denotative system, then the denotative action of *listen to* would be much more strongly supported, and would therefore have sufficient impetus to be implemented by the boss as a denotative action, the result of which will be to register the information provided by the *sponsor* and increase the denotative estimate of *C*'s competence.

This can either lead to a rational choice of this applicant over others, should the denotative distribution be sufficiently precise as a result of this, or can lead to a connotative choice of hiring or rejecting the applicant. As the applicant's connotative identity has now shifted to *competent applicant*, the deflections will be more significant in the decision. In this case, *boss hire competent applicant* has a deflection of 1.9, while *boss rebuff competent applicant* has a deflection of 3.1. It creates more social disruption to rebuff a competent applicant than a non-competent one, as expected.

9.1.3 Final network, Tradeoffs and Mutiny

After the employment of *C* by *boss*, *C* has become an *employee*, and so has gained the trust of the boss, as shown in Figure 27(b). The triadic relationship, originally one that is best supported by moderately high values of $\hat{\alpha}$ and $\hat{\gamma}$, with reduced $\hat{\delta}$, has moved to one that supports higher $\hat{\alpha}$ at the expense of reduced $\hat{\gamma}$ (see Figure 14(c)). This is a network that requires more trust, and thus agent *C* would do better by shifting his parameters to better align with the group. This is simply a matter of aligning his model of uncertainty up with that of the *boss*.

The precise nature of how these tradeoffs will work out is something that is negotiable, and highly context dependent. Thus, an agent may be using a different model of uncertainty in each context. We can associate this model with an institution, containing denotative and connotative meanings for many elements of relevance in that context. The various institutions in which an agent participates combine to form the agent's *self*, something that contains both a denotative *persona*, and a connotative *self-sentiment*. You may, should you choose, imagine the agent's self as another model of a person, albeit one with multiple identities that are useful in different situations. More details on the origins of these ideas can be found in (MacKinnon and Heise, 2010; Hoey and Schröder, 2015).

Upon hiring the employee, the *boss* may extend their trust to other relationships that *C* is involved in, such as towards other employees. That is, *boss* expects *C* to get along with the other employees. Suppose, however, that the employee decides that he wants someone else to

⁶²done with Indiana 2005 data

be the boss, say agent D ? Then, he might start to establish political control amongst the other employees, and would seek structural holes that he could exploit that would be likely to set one or more of the employees against the boss. For example, C could discover that the boss likes agent B but not agent D , and further that agent D and B like each other. Agent C can now exploit this structural hole by convincing B that D would be better as boss.

Suppose that D (C 's friend) emerges as a potential new leader. This could happen denotatively (D is observed to be more competent than *boss*), or connotatively, (more people like D than like the *boss*). In either case, the *boss* may be able to “clamp down” and establish much more order, by imposing regulations, for example. Or, agent D may take over from the *boss* and assume control. In either case, the shift of the group will be significant, and the new order that emerges will be dependent on a number of factors. If D attempts to make this mutiny, having the support of C and B will be very important, for the same reasons as discussed in the sponsor example (although in this case some links may be negative).

9.2 Conformity

Similarly to the dissonance case, modeled with *BayesAct* in (Hoey et al., 2020), conformity experiments (Asch and Guetzkow, 1951) can be explained in the same way, as they show a shift in a person’s denotative representation (of the correct answer) towards the representation of the group. In these experiments, a participant is placed in a group of (what they think are) equals, whereas in fact the rest of the group are complicit in the experiment. A question with a really obvious correct answer is posed, and the group members all say the correct answer is the wrong one. Participants in these experiments will be more likely to select the (obviously incorrect) answer. In the same way as the previous experiment, we imagine a denotative X that represents *right* and *wrong* (analogous to *good* and *bad* in the dissonance example). Then, a prior over X is that the answer chosen (the correct one initially) is *right* with high probability, $P(X = \textit{wrong}) = 0.1$. The prior over Y is again the identity of the subject.

We then update the model sequentially by multiplying the denotative posterior by a probability of observing another group member selecting the *right* answer of 0.95 (0.05 probability of selecting the wrong answer, as there may be some other cause for them selecting the wrong answer), and use this as the prior for a second update. This is the same as adding an observation Ω_x to the model, as above, which is generatively linked to X_b through an observation function $P(\Omega_x|X_b)$ which is $P(\Omega_x = \textit{right}|X_b = \text{“right answer chosen”}) = 0.95$. Observation of a single other participant selecting an answer multiplies the posterior over X with the observation function $P(\Omega_x|X)$, and this becomes the new prior over X . This is repeated N times to get $P(X) = \prod_j P(\Omega_x = j|X_b)P(X_b)$. The posterior over Y is used directly as the new prior over Y .

Figure 28(a) shows the results of repeating this process five times for⁶³ $\mu_y = 1.81$, $\sigma_y = 1.3$ and $\gamma = 1.0$. Clearly, as more and more group members select the wrong answer, the posterior over X increasingly favors the *wrong* answer (rising to $P'(X = \textit{wrong}) = 0.98$ after five iterations). The posterior over Y also decreases, with increasing weight put on the *wrong* mode, indicating that the interpretation has shifted from the selected answer being *right* to the selected answer

⁶³These are the mean and standard deviation of evaluation ratings of the identity “student” in the Georgia 2015 dataset. The Asch experiment participants were male college students in the 1950s, who were rated at that time as (EPA:{1.4, 0.9, 1.0}) (printouts obtained from Osgood’s original work, available at <http://www.indiana.edu/~socpsy/Atlas/>), but variances are not available.

Figure 28: Simulation of conformity. (a) The posteriors over X and Y shift towards the observed evidence (of the selected answer being incorrect), causing a re-interpretation of a *right* answer as something *wrong*. $S(P')$ is the entropy of $P'(X)$ and $P'(\text{wrong})$ is the posterior probability of $X = \text{wrong}$. (b) plot of the overall trends in entropy and posterior probability as a function of iteration.

being *wrong* (and so the other answer, actually the incorrect one, should be selected). The model thus predicts negative feelings for the participant's sense of self (Y), as they cope with the fact they apparently got the answer wrong.

9.3 Exogenous events

Now consider what happens when denotative precisions are suddenly decreased, e.g., by a catastrophic event or a sudden change in the diversity of the population due to migration. The BayesACT analysis predicts that connotative meanings will become more salient. However, as we have seen, there are two forms of connotative beliefs, one describing more static elements, and one describing more dynamic ones. Those populations or groups who already rely more on static meanings will be even further “pushed” towards these static meanings under a forced decrease in denotative precision. Similarly, those groups who rely more on dynamic meanings will be pushed towards those dynamic meanings more.

One can see this clearly by taking it to the limit, and considering a group using an infinitely imprecise static model. In this case, the dynamic connotative model must take on all the increase in uncertainty forced upon the denotative model. At the other end of the spectrum, a group with an infinitely imprecise dynamic model must increase the precision of its static model in order to account for increased denotative uncertainty. In between these two, groups are pushed towards their current positions, and we see that the process of increasing denotative uncertainty results in a distinct polarization of the status and legitimacy biases. Those who relied on status (at the expense of trust), do so more after a calamity, and will seek authoritarian figures with high status, but emotional resonance, regardless of the intentions of those persons. Those who relied

on trust, do so more (at the expense of status) after a calamity, and will seek more egalitarian leaders with equal status who are trusted to enable social structures that encourage legitimacy. Thus, the two groups are driven apart in their politics by an increase in uncertainty. This polarization is explored in more detail in (Hoey and Schöder, 2020).

In a more controlled setting, changes in the denotative complexity of the task, or the cost incurred in the task, are controlled for (Muthukrishna et al., 2016). In these experiments, there is no deception, so the measurement is about conformist transmission, which only happens if the individual goes beyond what you would expect based on the majority alone, otherwise it is called unbiased transmission. That is, if 60% of the participants are voting for one option, then if the individual has a 60% chance choose that option in an experiment with no deception, then a conformist would exceed that percentage (and the measurement would yield, say, 90%). Given a measurement of this percentage, then the more options to choose from, the excess would be larger, indicating that a consensus of only a few other participants on one option would be more significant if there are more options to choose from.

In such cases participants (who are WEIRD), respond with increased social learning as indicated by increased conformist transmission (the excess in the previous paragraph increases) when the task complexity or cost is increased. I can simulate this in the same manner as in Section 9.2 by simply increasing the entropy of the denotative prior, $P(X = \textit{wrong})$. The result is a figure that echoes the results from human experiments in Muthukrishna et al. (2016, Figure 3). As the denotative complexity increases (as the problem becomes harder to solve), participants increasingly make use of social information, relying more heavily on their connotative systems.

Figure 29 shows the same as Figure 28 for higher and lower values of $P(x)$, and Figure 30(a) shows the overall trend. As expected, when $P(X)$ is less precise ($-\log(P(x))$ is lower), it takes fewer iterations to start accepting social information. The mapping from $P(X)$ to the number of options presented in (Muthukrishna et al., 2016) is not well defined, however, so the fact the curve does not appear to “level off”⁶⁴ is not as important as the overall trend.

Finally, I can also reproduce something akin to Muthukrishna et al. (2016, Figure 5a), shown in Figure 30(b). This shows that, as $P(\Omega_x|X_b)$ is less precise, conformity decreases at a fixed number of iterations (participants giving social information) of 2, as in (Muthukrishna et al., 2016). This is as expected, since the information becomes less meaningful, and social meanings are discarded more readily. Note the curve in Figure 30(b) does not look linear, but this is mostly because of the most high fidelity setting of 0.99, which would be an expectation of only one person in a hundred getting the answer wrong. Muthukrishna et al. (2016) tell participants that 100% of the information is social (aka reliable), but this still leaves the uncertainty the participant has *in general* about people’s information, which may be much lower. That is, as in the last example, there is no immediately clear mapping between my setting of $P(\Omega_x|X_b)$ in my model and that of transmission fidelity in Muthukrishna et al. (2016), although one could probably be constructed from the data.

Changing the raw payoff given to participants is found by Muthukrishna et al. (2016) to have little effect as soon as it rises above \$0. I believe this makes sense, as this payoff is strictly *individual*. In order to make it social, they should give higher payoffs for going with the group majority, and being correct. This creates a social dilemma in which the group may win if they

⁶⁴Something (Muthukrishna et al., 2016) claim for their Figure 2 and 4(a), but I have some difficulty agreeing with the need for a quadratic fit.

Figure 29: Simulation of conformity. Identical to Figure 28(b) in all cases except for (a) $P(x) = 0.5$ instead of $P = 0.1$; (b) $P(x) = 0.001$. The curve shifts towards the y-axis as $P(x)$ is more uncertain, meaning it takes less evidence to convince the participant to go with the group. See Figure 30(a) for overall trends as a function of this parameter. Keeping $P(x) = 0.1$ and changing the observation probability $P(o|x)$ (is 0.95 in Figure 28) to (c) 0.90 and (d) (0.99). As transmission fidelity increases (as $P(\Omega_x|X_b)$ increases, social information is more meaningful and a shift takes place after fewer social bits of information. See Figure 30(b) for overall trends as a function of this parameter.

Figure 30: (a) Decreasing denotative complexity (decreasing probability of being wrong *a-priori*, so $-\log(P(x))$ increases) implies increasing number of alters needed giving consensus before a change is made. (b) Increasing transmission fidelity leads to greater influence of social information, and thus to greater conformity (percentage who would switch after 2 iterations, as in (Muthukrishna et al., 2016)).

all contribute, but will lose if anyone doesn't conform. Otherwise the experiment should be the same, in that the individual rewards should still be available and made salient.

9.4 Ambiguity

The difficulty for humans to handle ambiguity is demonstrated with remarkable breadth by the inability of established firms to handle disruptive technology (Christensen, 1997). In Chapter 7, Christensen distinguishes between “unknown” and “unknowable”

“Not only are the market applications for disruptive technologies *unknown* at the time of their development, they are *unknowable*.” Christensen (1997, p.143, emphasis in original)

He also points out that managers will prefer high risk ventures in which the risk is *known* over any venture in which the risk is not known (ambiguous outcomes), and that ambiguity calls for *learning* and *exploration* strategies. In *BayesAct* terms, a known risk carries known sentiment, and so presents a situation with high denotative uncertainty, but low connotative uncertainty. An unknown risk, on the other hand, presents a situation with high connotative uncertainty, and *unknown* denotative uncertainty. Such a situation *requires* learning and exploration in order to find out what the risk is. Simply put, managers don't know how the unknowable will *feel*, and so shy away from it, opting instead for something connotatively predictable, but denotative unpredictable, in which they know the strategy is very risky, but can estimate expected value, and can predict how it will *feel* to engage in it (e.g. bold and audacious). Even if the expected value is low, the emotional value may still be high enough to motivate the action. Managers need to embrace the ambiguity presented by disruptive technologies, using it as an opportunity to explore new markets and develop new products. Academic researchers are the ideal candidates

to carry out research into these ambiguities, for they are not constrained by the value networks of the corporations.

These basic elements in *BayesAct* are in line with experiments showing how humans tend to act more pro-socially (cooperate in a public goods game) in ambiguous situations (Vives and FeldmanHall, 2018). It is also in line with experiments showing how the effects of increased ambiguity may result in more political polarization (Bail et al., 2018; Talisse, 2019). As people are exposed to more heterogeneous discussion online from unknown strangers, their political views become more deeply entrenched. More homogeneous discussion leads to more muted political beliefs being expressed. This is explainable by noting that the more heterogeneous discussion leads to greater denotative uncertainty, leading to a more heavy reliance on connotative priors, which may have stereotypes embedded into them, for example. As discussion becomes more homogeneous, self-affirmation increases the weight placed on available evidence (see Section 7.13) leading to a posterior biased towards a more neutral solution, if we assume that the evidence is, in fact, independent of the arguments presented and the stereotypes held and the solution is rational and valid.

In the 2020 movie “The Social Dilemma”, hype is made about the ability to control users, and to sell this control to advertisers. The movie makes a valid point about the ability of social networking companies like Facebook to modify elements of the interfaces to affect users. These modifications can have many effects, including an increase in political polarization by showing people news feeds that fit their beliefs, causing them to become more convinced. However, this does not imply that these companies can build an accurate model of a person based on their affective responses, and that they understand the relationship between the responder and the modified elements. In making this claim, the movie makers are confusing induction and deduction and are refusing to acknowledge the ambiguity inherent in the problem. While one may deduce that an affective response arises from a change made, it is incorrect to claim this means one can infer or induce a person’s inner self from their affective responses. While these responses may have resulted, in part, from their reactions to the change made, it may also depend on many other factors that are context or person dependent, including other social networks the person is involved with. I will write this out as a simple Bayesian network that characterises the $P(\text{Response}|\text{Change}, \text{Person}, Z)$, where Z are all other factors that impact the Person’s decision and Response to the Change.

The scientific error in the movie is a common one in which correlation is mistaken for causality, deduction for induction. For example, J. Philippe Rushton noted a correlation between brain size and a race’s economic wealth, but claimed that brain size was the cause of the race’s success, leading to great controversy.⁶⁵ It is highly likely that economic success by a race is due to a large number of factors other than brain size, few of which are properly quantified.

Thus, since a Response is a causal effect of the Change and the Person, one can deduce that the Response arose because of the Change (as we know the Person and Z remained fixed, which may not be valid as the change may affect the person and their interactions with Z), and one can even measure the results of an A/B test to see which Change is more apt at delivering a Response (again, assuming Person remains fixed), but one cannot infer the Person given a particular Response to a Change because of the unknown other factors Z impacting the affective

⁶⁵See the famous debate about this at <https://www.cbc.ca/archives/entry/the-rushton-suzuki-debate>. Many of his articles have since been retracted, see for example <https://retractionwatch.com/2020/06/17/elsevier-journal-to-retract-2012-paper-widely-derided-as-racist/>.

Figure 31: Bayesian network showing a simple model of a social network user’s response to a change. There may be links between the three “parent” variables, Z , Person and Change, but I have not included these for simplicity.

response. That is, given the BN above, we would be asking the query (P =Person, R =Response, C =Change):

$$P(P|R, C) = \sum_Z P(P, Z|R, C) \propto \sum_Z P(R|P, C, Z)P(P)P(Z)$$

Even ignoring the prior over Z (which is unknown), simply *evaluating* this density requires summing over *all* possible other factors, clearly infeasible. In the words of Lévi-Strauss (1967, p.22), the social networking companies are ‘bricoleurs’ who define structure by means of observed events (outcomes, responses), but they claim this gives them enough information to be ‘scientists’ who define events by means of structures (e.g. theories).

9.5 Political Polarization

Political polarization arises in two different ways, one of which is described in (Bail et al., 2018), in which exposure to opposing viewpoints increases political polarization, and one depicted in the film “the social dilemma” in which non-exposure to opposing viewpoints (e.g. “echo chambers”) increases political polarization. I hypothesise that while the latter (non-exposure to opposing views, which I will label the “Hollywood” case) has a denotative or deliberative effect, the former (exposure to opposing views, which I will label the “Bail” case) has a connotative effect, both of which are to increase polarization. A similar argument is made by Facciani (2020), who states “*I argue that including the variable with whom the participant discusses ‘important matters’ is crucial for accounting for these inconsistencies.*” That is, whether the experiments ask for information about connotatively meaningful persons or not is important for gauging the resulting effect. Minimal social connection in the Bail experiment leads to the information being treated denotatively (and rejected, making the denotative posterior more dispersed), while in the Hollywood account information is treated connotatively, because it arises from trusted others, thus making the connotative prior more precise and strong). Essentially, when discussing a different viewpoint with a “trusted” alter, the connotative cost of “disagreeing with” this alter is high, so the act of agreement becomes much more salient and important, and the evidence trusted more. When discussing a different viewpoint with a stranger, the problem is denotative, so the evidence is compared using the individuals currently (stereotyped or biased) model.

In the Hollywood case, an agent’s denotative beliefs $P(x)$ will be shifted towards those x which are consistent with a particular political viewpoint, $\mathcal{S}$, that is, $x \in \mathcal{S} \rightarrow x \in \mathcal{S}_r$ (for republicans)

or $x \in \mathcal{S}_d$ (for democrats). Thus, democrats will have $P(x) = 0$ for $x \notin \mathcal{S}_d$, and republicans will have $P(x) = 0$ for $x \notin \mathcal{S}_r$. Simply, it is the polarization of information that causes the polarization of people. If someone who believes the earth is flat never sees “Earth Rise”,⁶⁶ then perhaps they will have a $P(\text{Earth_is_round}) = 0$ that never changes. If someone unsure about the shape of the earth sees only information claiming it’s flat, their beliefs will shift towards that part of the state space. If, prior to the polarization of information, republicans and democrats have some density over the opposing viewpoint (and vice-versa), then, after the information is polarized, the portion of this belief density on the states of the opposing viewpoint becomes weaker as it is not supported by evidence, and thus the two distributions by definition must move apart. Notice that the polarization of information is also expected to reduce the variance in these distributions, as the evidence has become more consistent.

On the other hand, in the Bail case, agents’ denotative beliefs are made *more* dispersed (less well localized) in the experimental condition. As I have discussed at length, this leads to an increased focus on the connotative state. Agents in this situation will be emotionally polarized, and will “go with their gut.” Suppose a person who votes republican believes that certain policies are more appropriate or beneficial. These policies, through the somatic transform, are considered culturally the “things republicans do,” such as decreasing taxes and increasing military spending. Thus, this person’s identity is well supported by these policies. Exposure to other policies, considered culturally the “things democrats do,” like increasing taxes and legalizing marijuana, will be in conflict with this person’s identity, and so must be discounted (because otherwise the only other options are to either shift their connotative beliefs about themselves, something difficult to do as it connects to the more slowly changing self, or to counteract it denotatively by turning the information content off, if possible). Their connotative beliefs therefore *strengthen* in the face of the disruption, in order to counteract the novel contradictory evidence. Consider a person who considers behaviours like “concede” with the opposing party because these make denotative sense (we all have to get along, otherwise we’ll end up in a civil war). Even though they do not believe in the other party’s ideals, they are willing to go along with it for the good of their own survival. Now, exposure to (especially emotionally inflammatory) contradictory denotative evidence, such as a change in the laws around abortion, leads to greater ambiguity denotatively, because the previous computation of the positive outcome from the behaviour “concede” is based on things staying as they are. The resulting shift towards a more precise connotative belief will “rule out” or weaken the behaviour “concede”, and strengthen the likelihood of behaviours such as “defeat” or “lambaste” (e.g. “Democratic Senator lambastes Republican President”). This is precisely because conceding to republican policies is something very unlike a democrat. Thus, the prior connotative beliefs about party affiliation strengthen, and so does the political polarization in the group.

The observation that people empathise more with others who have the same “moral matrix” is evidence for the same type of alignment process of the individual to the group (Haidt, 2012). The *moral matrix* in this case is simply the position on the freedom-equality axes, but is much more context dependent than Moral Foundations Theory (Haidt, 2012) and does not consider the intuitionist component to be necessarily heritable or stable and dispositional. This is in line with recent (lack of) evidence for the five-factor moral foundations model, pointing instead to a two dimensional model (Smith et al., 2017). Instead, *BayesAct* considers “intuitive reasoning”

⁶⁶A famous photograph of the Earth from space (<https://www.lpi.usra.edu/resources/apollo/frame/?AS08-14-2383>), which many say really instilled the roundness (and smallness) of the earth in their minds.

to be based somewhat on a long-term history of identities that are summed up in the *self*, combined with a situational identity chosen by the individual given the context. That is, *BayesAct* models shift their “moral matrix” according to the group or social context, and both group and individual do a situational “dance” in order to find an aligned matrix to work from. The result is a clear prescription for solving problems of political polarization, and it is not political, and is reasonably well known now (Talissee, 2019). It is simply that each person must be (1) confident in who they are, self-affirmed; and (2) truly empathetic with the people on the “other side”. This finding is reflected as well by (Facciani and Brashears, 2019), and the results of Bail et al. (2018) show that should (2) not hold, polarization increase as homogeneity increases.

BayesAct can be tied to economic and political theorizing by noting that different groups will be operating on different regions of the simplex. Governments have some degree of control, certainly over the structure of the denotative world, and over its degree of validity (and thus $\hat{\delta}$). One simple such control is over interest rates, and one aspect that it cannot so easily control is inflation (the “discount rate” see Section 4.2). Changing such a discount rate indirectly (i.e. by printing money) will lead to polarization of political extremes, since attempting to move “up” the simplex (by increasing $\hat{\delta}$ artificially) will cause reactions on both political sides of the $\widehat{A\Gamma}$ axis towards their respective centers ($\hat{\alpha}$ and $\hat{\gamma}$). This will cause increased separation of the groups on the simplex. Rather, governments that couple their control mechanisms with *narratives* that are emotional (Shiller, 2017) and provide some forcing towards one or the other of $\hat{\alpha}$ or $\hat{\gamma}$, will be more successful in maintaining the status quo while controlling for economic growth. Nevertheless, the emotional reactions of large political groups is also quite unpredictable, and may go through various *tipping points* at which the entire structure changes very rapidly, including uncontrolled changes to inflation rates. For example, a large portion of conservative-minded thinkers suddenly getting the idea that there is a conspiracy on, will retreat vigorously towards their “corner”, and create greater instability. The social engineering of such narratives and tipping points with specific goals in mind is a complex problem in its own right.

Finally, I make a note on the current political and social situation being experienced worldwide. I will focus on the United States as they have experienced multiple forces simultaneously. The 2020 pandemic caused an increase in isolationism, a shift away from globalization. This necessarily makes the world more *valid*, and thus forces the society “up” the simplex I have described. This leads to more conservative (illiberal) governance, and a stronger dependence on authority. It also brings with it increased focus on inequality. In order for functional static rules to be made, people must not shift in inequality. This increased focus erupted in the USA due to some key catalyst events, but strengthened by the focus on inequality caused by the pandemic. Thus, the natural reaction is to cause disruption and chaos, which many Americans did. This *decreases* validity and forces the society back “down” the simplex, albeit perhaps in a different location than before. Thus, I expect that depending on the motivation of the protesters, the move towards a more conservative government may be slowed.

However, as the method of choice by currently right-leaning governments is to enforce greater constraints on people, which enforces a social structure that is more conducive to conservative thinking, thereby reinforcing it. The proliferation of sets of overly complex rules in late October 2020 (e.g. exams at a local university with pages of densely packed rules, or the organization of Halloween trick-or-treating using highly complex sets of rules which accomplish almost nothing) demonstrates the stronghold taken by these conservative-leaning people who are really embracing their conservativeness. Liberal polarization is less severe, with more reliance on “trust” (e.g.

“we are all in this together”), for which the appropriate response is silence, patience, and repose, all of which do not serve to push the liberal agenda forwards, and leaves center stage for the sole possession of the conservative wing.

9.6 David Heise

In trying to piece together how the management of uncertainty was related to the parameters of *BayesAct*, I traveled to Indiana in the summer of 2018 to visit with David Heise. I showed him the tradeoff curve for perfectly invalid environments. In fact, this is to some degree what sociologists look for (invalidity) as they are modeling the behaviours of large groups of people. Therefore, any actions taken due to knowledge of the denotative world will show up as noise. In any case, this curve showed a simple downwards-going diagonal line connecting two orthogonal axes, going from α to γ (instead of $\hat{\alpha}$ and $\hat{\gamma}$ and missing $\hat{\delta}$).

Figure 32: (a) Drawing by David Heise showing imagined social groups at different locations on a $\alpha - \gamma$ space. The diagram would look as in (b) if plotted on the three dimensional simplex.

After initial skepticism followed by an understanding of the basic principle, he proceeded to label the diagram in terms of groups he might expect to see at that setting. His original labeling is shown in Figure 32(a), and is replicated in Figure 32(b) where it is rotated so it fits with my three dimensional scheme.

The groupings Heise suggested fit well with my interpretation of the parameters, and also with my expectations of what sorts of structures will arise at these parameter settings. The funeral group, for instance, is a highly valid environment, with fixed rules or norms. It however has less emotional overtones ($\hat{\alpha}$ and $\hat{\gamma}$), most of which is meant to be kept fairly reserved. A New York street scene, on the other hand, is highly invalid. The street is full of strangers of all types and

it is difficult to predict what will happen next. Again, however, there is little emotion as most people are not engaging with others. Further, the degree of diversity is high.

In the other direction, in Central Park one encounters not such an invalid environment as out on the street, as people are sitting, engaging in eating or playing. However, the degree of trust required by all parties is very high. That is, all must enjoy the park together, so for the time being, all are accepting of all others (i.e. “I’ll leave you that picnic spot if you leave me this one”). The degree of diversity is also high. Thus, we see that $\hat{\alpha}$ is high, validity ($\hat{\delta}$) is moderate and $\hat{\gamma}$ is low. On the other end of things is the poker game, which is a small clique of people, the environment is moderately invalid (hidden information). Based on that, I can infer that given the clique social structure, I expect to see high $\hat{\gamma}$ and low $\hat{\alpha}$. Another example of high $\hat{\gamma}$ is a so-called “barn raising” in which a community comes together to help a neighbour whose barn has fallen down. Some communities do this exceedingly quickly and efficiently, but maintain a highly homogeneous state in which all contributors are attempting to erect the same exact barn. More diverse communities may achieve the same thing through division of labour, requiring trust (higher $\hat{\alpha}$) in order that each contributor remains secure that the others will do their jobs if she does hers.

9.7 Automatic Government

If we agree that unity is better than divisiveness in society, then there will need to compromise. Doing it all one way for four years, and then switching to all the other way leaves no one satisfied. It is much more efficient if the two sides simply *decide* to do it in a *definable*, *interpretable* and *politically neutral* way. That is, for any decision to be made, an algorithm determines the most politically neutral course of action, and the society as a whole agrees to go that route through some direct democratic process (say a referendum, where a non-vote is a vote for the neutral point). Each and every political decision thus considered is available to each and every citizen and each can vote on the choice. The political decisions are explainable to each individual in an individualized way. That is, an illiterate person would not receive a portfolio of financial reports as would the economics professor, but would be presented with a simple-language version, or something they could listen to that would explain how this neutral solution works *in their context*.⁶⁷

The social contract of such a society would be one of agreeing to abide by the neutral political decision for each national level decision that needs to be made. It may be the case that each state would also implement a neutral-point deciding method, and each association within that state, and each work team within that association, all the way down. It may also be possible to *set* the neutral point to something that is not exactly balanced, based on the social constitution of the group being served. That is, at a state level in the United States, the neutral point may be set at a more conservative baseline in Texas than in Vermont. How to set the baseline then becomes a second democratic process, but could be done through simulations or questionnaires. That is, people could be asked to respond to sentiment questionnaires or situation questionnaires (e.g.

⁶⁷This opens up the option of truly allowing everyone to participate, even madmen and children. Each specialized group could be given options in a way that is sensical to them. Perhaps a toddler would need to be offered one of two dolls. This may be taking it too far. There might have to be an age cut off, and then a sanity cut off. However, all marginalized or disfavored groups would have the exact same opportunity, meaning there would have to be public voting booths everywhere, but these would look something like a phone booth. Online solutions would clearly also work should security be guaranteed.

“what would you do if...”) and these would be used to construct the theoretically ideal neutral point. In fact, the center point can be updated on an ongoing basis based on the outcomes of the referenda that are held regularly.

Usually, government attempts to handle a resource request, such as raising a tax, or tendering a contract, by evaluating how it will impact the freedoms of various groups of people. For example, a decrease in taxation may increase positive freedom for some individuals, while decreasing it for others. The savvy politician or political group will evaluate these freedom changes, and will apply some weighting function to make a final determination. While so-called *algorithmic government* may provide automated estimates and evaluations of these freedom changes (Engin and Treleaven, 2019), *automatic government* (what I am proposing) provides automated weighting of them to ensure a neutral political choice. The neutral point is set according to the sentiments of the whole people. That is, the connotative distribution of the entire population is measured, giving a sentiment distribution that can be used, along with the Bayesian model above, to compute expectations over possible actions to take. These expectations define the neutral point, and any decisions take into account that neutral point (but could average over a swatch of neutral points). It is not believed that the median point in this distribution will shift very fast, but re-adjustments to the neutral point are possible, and different neutral points may be used at different levels of organizational hierarchy (e.g. national vs. local government). If different neutral points are used at the state level, for example, then the national level becomes a process of weighting the different state neutral points. The proposal of three somewhat independent government functions (presidential, congressional, and judicial) is a proposal to generate the weighting using political power. However, this is open to exploitation by a despot, who can shift power in a direction of choice, for example, by only listening to the third of the citizens who voted for him.

Therefore, automatic government makes choices by simply picking the neutral point on the simplex. The main difference between a social democracy and automatic government is that instead of changing the state to match the people (which is often not possible to do in a polarized state), you change the people to match the state. Each person has to accept the neutral point decision, but can do so because automatic government gives them two very good reasons to. First, the decision is the “will of the people” as it reflects their sentiments. Second, no one is better off than anyone else. Someone may be worse off, but only because they hold more extreme views.

Automatic Government works as follows:

1. A resource request is processed. This could be an efficiency detected by an algorithmic analysis of work patterns in the government, or it could be a request from a foreign government for a loan.
2. The request is evaluated directly using measures of uncertainty on the three dimensions of freedom. That is, to measure $\hat{\gamma}$, we measure the variance in sentiment across the whole population. To measure $\hat{\alpha}$ we measure the variance in impressions of events. To measure $\hat{\delta}$, we present people with a variety of standard dilemmas or tests so their management of denotative uncertainty can be measured.
3. The request is then formulated in the *BayesAct* model so defined with a query of the form “what would happen should A to B to O?”

4. The resulting economic reward is weighted by the deflection through the somatic transform. That is, the reward function is averaged over futures according to the deflection, with lower deflection futures contributing more. Note here that there is no weighting function that determines which of these two factors has a bigger effect. The connotative solution in *BayesAct* simply “guides” the denotative search for optimal reward in a certain direction.
5. A decision-theoretic (reward maximizing) choice is then made based on that reward function.

Thus, the request can be evaluated as the deflection of the statement “what would happen should A to B to O?,” given a set of transformation functions (e.g. a dictionary and an impression formation equation). The choice of transformation function could be one that is aligned with one particular political group (say it could have high $\hat{\delta}$ but low $\hat{\alpha}$ and $\hat{\gamma}$), or it could be the theoretically neutral point at which $\hat{\alpha} = \hat{\gamma} = \hat{\delta}$. Depending on this alignment, the deflection may be lower or higher, essentially biasing any other decision making towards or away from, respectively, the stimulus. If the neutral point is chosen, then, by definition, no group is favored, and the decision maker is relieved of the burden of trying to rank the groups.

That is, if the neutral point is not automatically chosen (as it is normally not), then the social order would be forced to try to find a weighting for each of the social classes or political groups. Modern social democracies, such as in the United States, function by shifting this weighting back and forth every few years between more extreme points. Others, such as in Canada, also shift the weighting back and forth, but the pendulum swings much less far from the center.

There are three stages to setting up an automatic government. In the first, the people’s sentiment is measured, and a *BayesAct* template is created. In the second, resource requests are translated into *BayesAct* models, which are used to generate a decision and a confidence. The confidence is translated into the scale, N, of a referendum, which is held in the third stage, in which the decision is presented to the people for a vote. If the confidence is high, only a small sample of people is needed (the first N who vote, say). If the confidence is low, a larger sample is needed. Suppose a divisive issue such as gun rights is involved in a resource request. Then the *BayesAct* decision will weight the economic harms and benefits of each side (more guns, more business, more crime, etc) with the social harms and benefits of each side (gains and losses of freedom). Given that both of these are very balanced, the decision confidence will be very low. A large sample of votes is therefore needed to side one way or the other.

Resource requests come from people and must have a sufficient number of signatories to be considered for the three stages above. Let’s take an example in which some (sufficiently large) group requests a resetting of the neutral point. This will be evaluated by the model by using the best possible economic prediction function (which is generated by the administrative branch, perhaps one using algorithmic government), with the values of freedom changes on people’s future economic behaviour. If it improves things sufficiently, it is brought to a vote, with a confidence set by how big the improvement is. If the improvement is obvious, perhaps only a very small referendum is needed to make this momentous shift.

Note that in automatic government, there is no need for an executive branch (e.g. a president and cabinet), for it is replaced with the neutral point. However, there is still need for a legal system (which evaluates laws) and an administrative government (which administers and enforces the laws). The only thing that has changed is the political propensity of decisions, which is now agreed upon by all to be neutral. The neutral point is defined by the people, the set of which can

now be as inclusive as possible, allowing prisoners, children, and madmen, to propose resource allocations and to vote in referenda. Maybe some would have to be excluded, but this would only be for punitive reasons (i.e. “you can’t vote because we, as a society, have agreed to punish you in this way, not because we don’t want to listen to you”), not for its impact on automatic government.

10 Conclusion

I have presented a deep Bayesian model of some elements of intelligent reasoning, in particular one that combines connotative (emotional) and denotative (deliberative) reasoning in a principled way. I have demonstrated how this model can account for certain properties of small social networks (triads) that have been well described in the literature, but have lacked a consistent theoretical grounding. I further showed how this model associates with some key sociological theorizing, in particular from studies of cultural sociology. Finally, I have related the model to work in political and social science, artificial intelligence and computational social science. My objective throughout was to link the management of uncertainty to the formation of social structures, and I hope you, dear Reader, are at least somewhat intrigued by my arguments to this effect and are not too tired out after reading this very long essay.

Acknowledgments: THEMIS.COG is funded by the Canadian Natural Sciences and Engineering Research Council and Social Sciences and Humanities Research Council, by the National Endowment for the Humanities (NEH) and the National Science Foundation (NSF) in the USA, and by the Deutsche Forschungsgemeinschaft (DFG) in Germany.

References

- K. J. Åström. Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. App.*, 10:174—205, 1965.
- George A. Akerlof and Rachel E. Kranton. Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753, 2000. doi: 10.1162/003355300554881. URL <http://qje.oxfordjournals.org/content/115/3/715.abstract>.
- Robert R. Alford. *Party and Society: The Anglo-American Democracies*. Greenwood Press, Westport, Conn., 1963.
- Jens Ambrasat, Christian von Scheve, Gesche Schauenburg, Markus Conrad, and Tobias Schröder. Unpacking the habitus: Meaning making across lifestyles. *Sociological Forum*, 31(4):994–1017, 2016. doi: 10.1111/socf.12293. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/socf.12293>.
- Elizabeth Anderson. *Private Government*. Princeton University Press, 2017.
- Rudolf Arnheim. *Art and Visual Perception*. University of California Press, 1954.
- Solomon E Asch and Harold Guetzkow. Effects of group pressure upon the modification and distortion of judgments. *Documents of gestalt psychology*, pages 222–236, 1951.

- Nabiha Asghar and Jesse Hoey. Monte-Carlo planning for socially aligned agents using Bayesian affect control theory. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, pages 72–81, 2015.
- Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563: 59–64, october 2018.
- R Axelrod and WD Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.
- Paul B. Badcock, Karl J. Friston, and Maxwell J.D. Ramstead. The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*, 31:104 – 121, 2019. ISSN 1571-0645. doi: <https://doi.org/10.1016/j.plrev.2018.10.002>. URL <http://www.sciencedirect.com/science/article/pii/S1571064519300028>. Physics of Mind.
- Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1804840115. URL <https://www.pnas.org/content/115/37/9216>.
- Robert Freed Bales. *Social Interaction Systems: Theory and Measurement*. Transaction Publishers, New Brunswick, NJ, 1999.
- Robert Freed Bales. Symlog website. <https://www.symlog.com/UsingSymlog/SYMLOGInDepth.aspx>, July 2020.
- Lisa Feldman Barrett and Ajay Satpute. Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain. *Current Opinion in Neurobiology*, 23:361–372, 2013.
- Richard E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.
- Peter L. Berger and Thomas Luckmann. *The Social Construction of Reality*. Doubleday: Anchor Books, 1967.
- Isaiah Berlin. *Liberty*. Oxford University Press, 2002.
- Elisa Jayne Bienenstock, Phillip Bonacich, and Melvin Oliver. The effect of network density and homogeneity on attitude polarization. *Social Networks*, 12(2):153–172, 1990.
- Pierre Bourdieu. *The Logic of Practice*. Stanford University Press, 1990.
- Craig Boutilier, Thomas Dean, and Steve Hanks. Decision theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.
- Ronen I. Brafman and Moshe Tennenholtz. R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.
- Valerie Braithwaite. Beyond rokeach’s equality-freedom model: Two-dimensional values in a one-dimensional world. *Journal of Social Issues*, 50:67–94, 1994.

- Joost Broekens, Elmer Jacobs, and Catholijn M. Jonker. A reinforcement learning model of joy, distress, hope and fear. *Connection Science*, 27:215–233, 2015.
- Jelle Bruineberg and Erik Rietveld. What’s inside your head once you’ve figured out what your head’s inside of. *Ecological Psychology*, 31(3):198–217, 2019. doi: 10.1080/10407413.2019.1615204.
- Ronald S. Burt. Network items and the general social survey. *Social Networks*, 6(4):293 – 339, 1984. ISSN 0378-8733. doi: [https://doi.org/10.1016/0378-8733\(84\)90007-8](https://doi.org/10.1016/0378-8733(84)90007-8). URL <http://www.sciencedirect.com/science/article/pii/0378873384900078>.
- Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.
- Aitor Calo-Blanco, Jaromír Kovářík, Friederike Mengel, and José Gabriel Romero. Natural disasters and indicators of social cohesion. *PloS one*, 12(6):e0176885, 2017.
- Nuttapong Chentanez, Andrew G. Barto, and Satinder P. Singh. Intrinsically motivated reinforcement learning. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1281–1288. MIT Press, 2005.
- Clayton M. Christensen. *The Innovator’s Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business Review Press, 1997.
- Claudio Cioffi-Revilla. *Introduction to Computational Social Science: Principles and Applications*. Springer, 2014.
- Geoffrey L. Cohen, J. ARonson, and C.M. Steele. When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin*, 26:1151–1164, 2000.
- Auguste Comte. *Cours de Philosophie Positive*. Bachelier, Paris, 1835.
- Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970. doi: 10.1080/00207727008920220. URL <https://doi.org/10.1080/00207727008920220>.
- Axel Constant, Maxwell JD Ramstead, Samuel PL Veissiere, John O Campbell, and Karl J Friston. A variational approach to niche construction. *Journal of The Royal Society Interface*, 15(141):20170685, 2018.
- Charles Horton Cooley. *Human Nature and the Social Order*. Scribner’s, New York, 1902.
- Matthew Corritore, Amir Goldberg, and Sameer B. Srivastava. The new analytics of culture. *Harvard Business Review*, 98(1):76–83, 2020. URL <https://escholarship.org/uc/item/8n75q20g>.
- Francis Crick. Function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Sciences*, 81:4568–4590, 1984.
- Antonio R. Damasio. *Descartes’ error: Emotion, reason, and the human brain*. Putnam’s sons, 1994.

- Roy G. D'Andrade. Cultural meaning systems. In *Culture Theory: Essays on Mind, Self and Emotion*, chapter 3, pages 88–122. Cambridge University Press, 1984.
- Adnan Darwiche. Human-level intelligence or animal-like abilities? *Communications of the ACM*, 61(10):56–67, October 2018. doi: 10.1145/3271625.
- Guillaume Deffuant, Sylvie Huet, and Frédéric Amblard. An individual-based model of innovation diffusion mixing social value and individual benefit. *American Journal of Sociology*, 110(4):1041–1069, January 2005.
- Daniel Defoe. *Robinson Crusoe*. William Taylor, 1719.
- René Descartes. *Discourse on Method*. 1637.
- Finale Doshi-Velez. The infinite partially observable Markov decision process. In *Advances in neural information processing systems*, pages 477–485, 2009.
- Michael O. Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts Amherst, January 2002.
- R.I.M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469 – 493, 1992. ISSN 0047-2484. doi: [http://dx.doi.org/10.1016/0047-2484\(92\)90081-J](http://dx.doi.org/10.1016/0047-2484(92)90081-J). URL <http://www.sciencedirect.com/science/article/pii/004724849290081J>.
- Robin I. Dunbar. From the keynote address at IC2S2 2015 in Helsinki Finland, June 2015.
- Emile Durkheim. *The Division of Labor in Society*. Free Press, 2014/1893.
- Magy Seif El-Nasr, John Yen, and Thomas R. Iorger. FLAME - fuzzy logic adaptive model of emotions. *Autonomous Agents and Multiagent Systems*, 3:219–257, 2000.
- Kevin Elliott and Daniel McKaughan. Nonepistemic values and the multiple goals of science. *Philosophy of Science*, 81(1):1–21, 2014. doi: 10.1086/674345.
- Richard M. Emerson. Social exchange theory. In Morris Rosenberg and Ralph H. Turner, editors, *Social Psychology: Sociological Perspectives*. Basic Books, New York, 1981.
- Zeynep Engin and Philip Treleaven. Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies. *The Computer Journal*, 62(3): 448–460, 03 2019. ISSN 0010-4620. doi: 10.1093/comjnl/bxy082. URL <https://doi.org/10.1093/comjnl/bxy082>.
- Robert Erikson and John H. Goldthorpe. Has social mobility in britain decreased? reconciling divergent findings on income and class mobility. *The British Journal of Sociology*, 61:211–230, 2010.
- Geoffrey Evans and Noah Carl. Class and politics. In William Outhwaite and Stephen P. Turner, editors, *The Sage Handbook of Political Sociology*, chapter 5. Sage Publications, 2018.
- Geoffrey Evans, Anthony Heath, and Mansur Lalljee. Measuring left-right and libertarian-authoritarian values in the british electorate. *British Journal of Sociology*, 47(1), March 1996.

- H.J. Eysenck. *The psychology of politics*. Routledge, London, 1954.
- Matthew Facciani and Matthew E. Brashears. Sacred alters: The effects of ego network structure on religious and political beliefs. *Socius: Sociological Research for a Dynamic World*, 6:1–16, 2019.
- M.J. Facciani. *How Self-Sentiments and Personal Networks Impact Political Polarization*. PhD thesis, University of South Carolina, Columbia, SC, 2020.
- Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999. URL doi:10.1162/003355399556151.
- Scott L. Feld and Richard Elmore. Patterns of sociometric choices: Transitivity reconsidered. *Social Psychology Quarterly*, 45(2):77–85, 1982.
- Oriel FeldmanHall and Amitai Shenhav. Resolving uncertainty in a social world. *Nature Human Behaviour*, 3:426–435, April 2019.
- Leon Festinger. Cognitive dissonance. *Scientific American*, 207(4):93–107, 1962.
- Benjamin Fish and Luke Stark. Reflexive design for fairness and other human values in formal models, 2020.
- Johnny R. J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18:1050 – 1057, 2007.
- Robert Freeland and Jesse Hoey. The structure of deference: Modeling occupational status using affect control theory. *American Sociological Review*, 83(2), April 2018.
- Noah E. Friedkin. Social cohesion. *Annual Review of Sociology*, 30:409–425, 2004.
- Milton Friedman. *Capitalism and Freedom*. University of Chicago Press, 1962.
- Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11((2)):127–138, 2010.
- Karl Friston. A free energy principle for a particular physics, 2019.
- Karl Friston, S. Samothrakis, and R. Montague. Active inference and agency: optimal control without cost functions. *Biological Cybernetics*, 106(8-9):523–541, 2012.
- E. Fromm. *Man for himself: An enquiry into the psychology of ethics*. Routledge & Kegan Paul, London, 1949.
- Francis Fukuyama. *Trust: The Social Virtues and the Creation of Prosperity*. Free Press, 1995.
- Anthony Giddens. *Modernity and Self-Identity*. Stanford University Press, 1991.
- Christopher M. Glaze, Alexandre L.S. Filipowicz, Joseph W. Kable, Vijay Balasubramanian, and Joshua I. Gold. A bias-variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nature Human Behaviour*, 2:213–224, March 2018.

- Amir Goldberg and Sarah K. Stein. Beyond "social cognition": Associative diffusion and the emergence of cultural variation. *American Sociological Review*, 53(5):897–932, 2018.
- Harry F. Gollob. The subject-verb-object approach to social cognition. *Psychological Review*, 81:286–321, 1974.
- P. Gomez Esteban and D. Rios Insua. An affective model for a non-expensive utility-based decision agent. *IEEE Transactions on Affective Computing*, pages 1–1, 2017. ISSN 1949-3045. doi: 10.1109/TAFFC.2017.2737979.
- Mark Granovetter. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3):481–510, Nov. 1985.
- Jonathan Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108:814–834, 2001.
- Jonathan Haidt. *The Righteous Mind: Why good people are divided by politics and religion*. Vintage, New York, 2012.
- Michael T. Hannan and John Freeman. The population ecology of organizations. *American Journal of Sociology*, 82(5):929–964, March 1977.
- N. Haslam, S. Loughnan, Y. Kashima, and P. Bain. Attributing and denying humanness to others. *European Review of Social Psychology*, 19(1):55–85, 2008.
- Joseph Heinrich. *The WEIRDest people in the world*. Farrar, Strauss and Giroux, New York, 2016.
- David R. Heise. *Expressive Order: Confirming Sentiments in Social Actions*. Springer, 2007.
- David R. Heise. *Surveying Cultures: Discovering Shared Conceptions and Sentiments*. Wiley, 2010.
- David R. Heise. Modeling interactions in small groups. *Social Psychology Quarterly*, 76:52–72, 2013.
- Casper Hesp. Hedging your bets: and active inference formulation of valence and arousal. Unpublished reserach project, 2018.
- Casper Hesp, Ryan Smith, Micah Allen, Karl Friston, and Maxwell Ramstead. Deeply felt affect: The emergence of valence in deep active inference, Dec 2019. URL psyarxiv.com/62pfd.
- Alain Van Hiel, Jonas De keersmaecker, Emma Onraet adn Tessa Haesevoets, Arne Roets, and Johnny R. J. Fontaine. The relationship between emotional abilities and right-wing and prejudiced attitudes. *Emotion*, 19(5):917–922, 2019.
- Jacob B. Hirsh, Raymond A. Mar, and Jordan B. Peterson. Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological Review*, 119(2):304–320, 2012.
- Steven Hitlin. *Moral Selves, Evil Selves: The Social Psychology of Conscience*. Palgrave Macmillan, New York, 2008.

- Hobbes. *Leviathan*. 1651.
- Jesse Hoey and Tobias Schröder. Disruption of status orders in societal transitions as affective control of uncertainty. *under review*, 2020.
- Jesse Hoey and Tobias Schröder. Bayesian affect control theory of self. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 529–536, 2015.
- Jesse Hoey, Tobias Schröder, and Areej Alhothali. Affect control processes: Intelligent affective interaction using a partially observable Markov decision process. *Artificial Intelligence*, 230: 134–172, January 2016.
- Jesse Hoey, Tobias Schröder, Jonathan Morgan, Kimberly B Rogers, Deepak Rishi, and Meiyappan Nagappan. Artificial intelligence and social simulation: Studying group dynamics on a massive scale. *Small Group Research*, 49(6):647–683, 2018.
- Jesse Hoey, Neil MacKinnon, and Tobias Schröder. Denotative and connotative control of uncertainty: A computational dual-process model. *under review*, 2020.
- Eric Hogewoning, Joost Broekens, Jeroen Eggermont, and Ernst G.P. Bovenkamp. Strategies for affect-controlled action-selection in soar-rl. In J. Mira and J.R. Álvarez, editors, *IWINAC*, volume 4528 Part II of *LNCS*, pages 501–510, 2007.
- Michael A. Hogg. Social identity theory. In Peter J. Burke, editor, *Contemporary Social Psychological Theories*, chapter 6, pages 111–136. Stanford University Press, 2006.
- Paul W. Holland and Samuel Leinhardt. A method for detecting structure in sociometric data. *American Journal of Sociology*, November 1970.
- Jeffrey R. Huntsinger, Linda M. Isbell, and Gerald L. Clore. The affective control of thought: Malleable, not fixed. *Psychological Review*, 121(4):600–618, 2014.
- R. Inglehart. *The Silent Revolution*. Princeton University Press, New Jersey, 1977.
- Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3040–3049, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL <http://proceedings.mlr.press/v97/jaques19a.html>.
- Kate Jeffery, Robert Pollack, and Carlo Rovelli. On the statistical mechanics of life: Schrödinger revisited. *Entropy*, 21(12):1211, Dec 2019. ISSN 1099-4300. doi: 10.3390/e21121211. URL <http://dx.doi.org/10.3390/e21121211>.
- Mateus Joffily and Giorgio Coricelli. Emotional valence and the free-energy principle. *PLoS Computational Biology*, 9(6):e1003094, 2013.
- Daniel Kahneman. *Thinking, Fast and Slow*. Doubleday, 2011.
- Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6):515–526, September 2009.

- Emmanuel Kant. *Groundwork of the Metaphysic of Morals*. 1785.
- Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *Proceedings of European Conference on Machine Learning*, 2006.
- Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, July 2009.
- Richard D. Lane. Neural correlates of conscious emotional experience. In Richard D. Lane and Lynn Nadel, editors, *Cognitive Neuroscience of Emotion*, chapter 15, pages 345–370. Oxford University Press, 2000.
- Edward J. Lawler, Shane R. Thye, and Jeongkoo Yoon. *Social Commitments in a Depersonalized World*. Russell Sage Foundation, 2009.
- David Lazer and Allan Friedman. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4):667–694, 2007. ISSN 00018392. URL <http://www.jstor.org/stable/20109947>.
- Claude Lévi-Strauss. *The Savage Mind*. The University of Chicago Press, 1967.
- Kurt Lewin. *Principles of Topological Psychology*. McGraw Hill, New York, 1936.
- Linda C. Li, Jeremy M. Grimshaw, Camilla Nielsen, Maria Judd, Peter C. Coyte, and Ian D. Graham. Evolution of wenger’s concept of community of practice. *Implementation Science*, 4(1):11, Mar 2009. ISSN 1748-5908. doi: 10.1186/1748-5908-4-11. URL <https://doi.org/10.1186/1748-5908-4-11>.
- Omar Lizardo and Michael Strand. Skills, toolkits, context and institutions: Clarifying the relationship between different approaches to cognition in cultural sociology. *Poetics*, 38:204–227, 2010.
- John Locke. *Essay Concerning Human Understanding*. Oxford University Press (2008), 1690.
- G. Loewenstein and J.S. Lerner. The role of affect in decision making. In R.J. Davidson, K.R. Sherer, and H.H. Goldsmith, editors, *Handbook of Affective Sciences*, page 619–642. Oxford Univ. Press, 2003.
- Charles G. Lord and Mark R. Lepper. Attitude representation theory. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 31, pages 265 – 343. Academic Press, 1999. doi: [https://doi.org/10.1016/S0065-2601\(08\)60275-0](https://doi.org/10.1016/S0065-2601(08)60275-0). URL <http://www.sciencedirect.com/science/article/pii/S0065260108602750>.
- Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109, 1979.
- Anna Lührmann, Juraj Medzihorsky, Garry Hindle, and Staffan I. Lindberg. New global data on political parties: V-party. V-Dem Institute Briefing Paper #9, October 2020.
- David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

- N. J. MacKinnon. *Symbolic Interactionism as Affect Control*. State University of New York Press, Albany, 1994.
- Neil J. MacKinnon and David R. Heise. *Self, identity and social institutions*. Palgrave and Macmillan, New York, NY, 2010.
- Neil J. MacKinnon and Jesse Hoey. On the inextricability and complementarity of cognition and affect: A review and model. *Under Review*, 2020.
- Robert P. Marinier III and John E. Laird. Emotion-driven reinforcement learning. In *Proc. of 30th Annual Meeting of the Cognitive Science Society*, pages 115–120, Washington, D.C., 2008.
- Peter V. Marsden. Core discussion networks of americans. *American Sociological Review*, 52(1): 122–131, February 1987.
- John Levi Martin. *Social Structures*. Princeton University Press, 2009.
- John Levi Martin. Life’s a beach, but you’re an ant, and other unwanted news for the sociology of culture. *Poetics*, 38:228–243, 2010.
- John Levi Martin and Alessandra Lembo. On the other side of values. *American Journal of Sociology*, 156(1):52–98, July 2020.
- Karl Marx. *Das Kapital. Kritik der politischen Ökonomie*. Verlag von Otto Meisner, Germany, 1867.
- Karl Marx and Freiderich Engels. *The Communist Manifesto*. Penguin Books, 1967 (1872).
- Robert R. McCrae and Paul T. Costa Jr. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of Personality*, 57(1):17–40, 1989. doi: 10.1111/j.1467-6494.1989.tb00759.x.
- Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.
- George Herbert Mead. *Mind, Self and Society*. University of Chicago Press, 1934.
- Hugo Mercier and Sperber. *The Enigma of Reason*. Harvard University Press, 2017.
- Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. Emotion in reinforcement learning agents and robots: A survey. *Machine Learning*, 107(2):443–480, 2017.
- J. L. Moreno. Contributions of sociometry to research methodology in sociology. *American Sociological Review*, 12(3):287–292, June 1947. URL <https://www.jstor.org/stable/2086518>.
- J.L. Moreno. *Who Shall Survive?: A New Approach to the problem of Human Interrelations*. Nervous & Mental Disease Publishing Co., Washington, D.C., 1934.
- Jonathan H. Morgan, Jun Zhao, Andrea Sedlacek, Lena Chen, Hayley Piper, Yliana Beck, Kimberly B. Rogers, Jesse Hoey, and Tobias Schröder. Modeling the culture of online collaborative groups with affect control theory, September 2019.

- Masahiro Mori. Bukimi no tani (the uncanny valley). *Energy*, 7(4):33–35, 1970.
- Michael Moutoussis, Pasco Fearon, Wael El-Deredy, Raymond J. Dolan, and Karl J. Friston. Bayesian inferences about the self (and others): a review. *Consciousness and Cognition*, 25: 67–76, 2014a.
- Michael Moutoussis, Nelson J. Trujillo-Barreto, Wael El-Deredy, Raymond J. Dolan, and Karl J. Friston. A formal model of interpersonal inference. *Frontiers in Human Neuroscience*, 8(160), 2014b.
- Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2021. URL <http://mlbayes.ai>.
- Michael Muthukrishna and Joseph Heinrich. Innovation in the collective brain. *Philosophical Transactions B*, 371(20150192), 2016.
- Michael Muthukrishna, Thomas JH Morgan, and Joseph Henrich. The when and who of social learning and conformist transmission. *Evolution and Human Behavior*, 37(1):10–20, 2016.
- Isabel Briggs Myers and Peter B Myers. *Gifts differing: Understanding personality type*. Nicholas Brealey, 2010.
- M. A. Nowak. Five rules for the evolution of cooperation. *Science*, 314:1560–1563, 2006.
- George Orwell. *Animal Farm*. Secker and Warburg, London, England, August 1945.
- Charles E. Osgood. On the Whys and Wherefores of EPA. *Journal of Personality and Social Psychology*, 12:194–199, 1969.
- Charles E. Osgood, G. J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, Urbana, 1957.
- Charles E. Osgood, William H. May, and Murray S. Miron. *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press, 1975.
- Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Canto Classics. Cambridge University Press, 2015. doi: 10.1017/CBO9781316423936.
- Orlando Patterson. Making sense of culture. *Annual Review of Sociology*, 40:1–30, 2014.
- Joseph M Paxton, Leo Ungar, and Joshua D Greene. Reflection and reasoning in moral judgment. *Cognitive science*, 36(1):163–177, 2012.
- Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
- Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.
- David Pitcher and Leslie G. Ungerleider. Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, xx(xx), Month 2020.
- Robert Plutchik. *The Emotions*. University Press of America, 1980.

- David Poole and Alan Mackworth. *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, 2010.
- Brittany Postnikoff. Robot social engineering. Master’s thesis, School of Computer Science, University of Waterloo, June, 2020.
- William T. Powers. *Behavior: The control of perception*. Aldine publishing co., Chicago, 1973.
- Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, NY., 1994.
- Matthew Rabin. A theory of fairness, competition and cooperation. *The American Economic Review*, 83(5):1281–1302, 1993.
- Maxwell JD Ramstead, Michael D Kirchhoff, and Karl J Friston. A tale of two densities: active inference is enactive inference. *Adaptive Behavior*, 0(0):1–15, 2019. doi: 10.1177/1059712319862774.
- Debajyoti Ray, Brooks King-Casas, P. Read Montague, and Peter Dayan. Bayesian model of behaviour in economic games. In *Proceedings of Neural Information Processing Systems*, 2008.
- Cecilia L. Ridgeway. *Status: Why is it everywhere? Why does it matter?* Russell Sage, 2019.
- M. Rokeach. *The nature of human values*. Free Press, New York, 1973.
- Jean-Jacques Rousseau. *The Social Contract*. 1762.
- Bertrand Russell. *A History of Western Philosophy*. Simon and Schuster, Inc., 1945.
- James A. Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.
- Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, third edition, 2010.
- Liam Paul Satchell, Roope Oskari Kaaronen, and Robert D. Latzman. The ecological approach to personality: Psychological traits as drivers and consequences of active perception. <https://doi.org/10.31234/osf.io/aqv6n>, 2020. URL <https://psyarxiv.com/aqv6n/>.
- Tobias Schröder, Jesse Hoey, and Kimberly B. Rogers. Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *American Sociological Review*, 81(4), 2016.
- Shalom Schwartz. Are there universal aspects in the structure and content of human values? *Journal of Social Issues*, 50:19–45, 1994.
- Pablo Sequeira, Francisco S. Melo, and Ana Paiva. Learning by appraising: an emotion-based approach to intrinsic reward design. *Adaptive Behaviour*, 22(5):330–349, 2014.
- Pedro Sequeira, Francisco S. Melo, Rui Prada, and Ana Paiva. Emerging social awareness: Exploring intrinsic motivation in multiagent learning. In *Proceedings of the 1st international joint conference on development and learning in epigenetic robotics*, pages 1–6, 2011.

- Daniel B. Shank and Rohan Lulham. Products as affective modifiers of identities. *Sociological Perspectives*, 2016.
- David K Sherman and Geoffrey L Cohen. The psychology of self-defense: Self-affirmation theory. *Advances in experimental social psychology*, 38:183–242, 2006.
- Robert J Shiller. Narrative economics. *American Economic Review*, 107(4):967–1004, 2017.
- Richard A. Shweder. In defense of moral realism: Reply to gabennesch. *Child Development*, 61: 2060–2070, 1990.
- David Silver and Joel Veness. Monte-carlo planning in large POMDPs. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS) 23*, pages 2164–2172. Curran Associates, Inc., 2010. URL <http://papers.nips.cc/paper/4031-monte-carlo-planning-in-large-pomdps.pdf>.
- David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.
- Herbert A. Simon. Motivational and emotional controls of cognition. *Psychological Review*, 74: 29–39, 1967.
- Mario Luis Small, David J. Harding, and Michèle Lamont. Reconsidering culture and poverty. *Annals of the American Academy of Political and Social Science*, 629:6–27, May 2010.
- Adam Smith. *The Theory of Moral Sentiments*. W.Strahan, London, 1759.
- Kevin B. Smith, John R. Alford, John R. Hibbing, Nicholas G. Martin, and Peter K. Hatemi. Intuitive ethics and political orientations: Testing moral foundations as a theory of political ideology. *American Journal of Political Science*, 61(2):424–437, April 2017.
- Ryan Smith, William D.S. Killgore, and Richard D. Lane. The structure of emotional experience and its relation to trait emotional awareness: A theoretical review. *Emotion*, 18(5):670–692, 2018.
- Ryan Smith, Thomas Parr, and Karl J. Friston. Simulating emotions: An active inference model of emotional state inference and emotion concept learning. *bioRxiv*, 2019. doi: 10.1101/640813. URL <https://www.biorxiv.org/content/early/2019/05/21/640813>.
- Lynn Smith-Lovin, Dawn T. Robinson, Bryan C. Cannon, Jesse K. Clark, Robert Freeland, Jonathan H. Morgan, and Kimberly B. Rogers. Mean affective ratings of 929 identities, 814 behaviors, and 660 modifiers by university of georgia and duke university undergraduates and by community members in durham, nc, in 2012-2014. University of Georgia: Distributed at UGA Affect Control Theory Website: <http://research.franklin.uga.edu/act/>, 2016.
- Luke Stark and Jesse Hoey. The ethics of emotion in AI systems. In *Proc. FaCCT**, 2020. doi: 10.31219/osf.io/9ad4u.
- Niklas K. Steffens, S. Alexander Haslam, Kim Peters, and John Quiggin. Identity economics meets identity leadership; exploring the consequences of elevated CEO pay. *The Leadership Quarterly*, 31, 2020.

- Henri Tajfel and John C. Turner. The social identity theory of intergroup behavior. In Stephen Worchel and William Austin, editors, *Psychology of intergroup relations*. Nelson-Hall, Chicago, 1986.
- Robert B. Talisse. *Overdoing Democracy: Why we must put politics in its place*. Oxford University Press, 2019.
- Paul Thagard and Tobias Schröder. Emotions as semantic pointers: Constructive neural mechanisms. In L. F. Barrett and J. Russell, editors, *The psychological construction of emotion*. Guilford, New York, 2014.
- Charles Tilly. *Durable Inequality*. University of California Press, Berkeley, CA, 1998.
- Hideki Toya and Mark Skidmore. Do natural disasters enhance societal trust? *Kyklos*, 67(2): 255–279, 2014.
- Johnathan H. Turner. The evolutionary biology and sociology of social order. In Edward J. Lawler, Shane R. Thye, and Jeongkoo Yoon, editors, *Order on the Edge of Chaos*, chapter 2, pages 18–42. Cambridge University Press, 2016.
- Stephen Vaisey and Lauren Valentino. Culture and choice: Toward integrating cultural sociology with the judgment and decision-making sciences. *Poetics*, 68:131 – 143, 2018. ISSN 0304-422X. doi: <https://doi.org/10.1016/j.poetic.2018.03.002>. URL <http://www.sciencedirect.com/science/article/pii/S0304422X17301869>.
- Kees van den Bos. Uncertainty management: The influence of uncertainty salience on reactions to perceived procedural fairness. *Journal of Personality and Social Psychology*, 80(6):931–941, 2001.
- Kees van den Bos and J. Miedema. Toward understanding why fairness matters: The influence of mortality salience on reactions to procedural fairness. *Journal of Personality and Social Psychology*, 79:355–366, 2000.
- Rufin VanRullen and Ryota Kanai. Deep learning and the global workspace theory. *under review*, 2021.
- Marc Lluís Vives and Oriel FeldmanHall. Tolerance to ambiguous uncertainty predicts prosocial behaviour. *Nature communications*, 9(2156), 2018.
- John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 3 edition, 1953.
- Max Weber. *The Theory of Social and Economic Organization*. 1915.
- Michael Weisberg. Three kinds of idealization. *Journal of Philosophy*, 104:639–659, 2007.
- Etienne Wenger. *Communities of Practice: Learning, Meaning and Identity*. Cambridge University Press, 1998.
- Harrison White. *Identity and Control*. Princeton University Press, Princeton, NJ, 1 edition, 1992.

Harrison White. *Identity and Control*. Princeton University Press, Princeton, NJ, 2 edition, 2009.

Harrison C White. Where do markets come from? *American journal of sociology*, 87(3):517–547, 1981.

Wilhelm Wundt. *Outlines of Psychology*. Wilhelm Englemann, 1896.