Cohesion and the Explanation of Constitutional Choice in Self-Governing Systems

Asimina Mertzani[1], Jeremy Pitt[1], Stefan Sarkadi[2], Madalina Sas[1], Matt Scott[1] and Ciske Smit[1]

[1]Department of Electrical & Electronic Engineering, Imperial College London
[2]Department of Informatics, King's College London, UK
j.pitt@imperial.ac.uk

Abstract

Self-governing social systems have to address a critical problem of cohesion. Cohesion is directly affected by constitutional choice, which is concerned with the nature and style of a political regime that produces or promotes qualitative human values such as safety, affinity and dignity. Evidence from the social sciences indicates that effective constitutional choice requires meaningful and justifiable explanations for a change in political regime, e.g. for congruence with prevailing environmental conditions or to encourage continued voluntary association. To investigate the relationship between explanation of constitutional choice and the emergence of cohesion, we formalise a regime change game in the AATS+V (Action Based Alternating Transition Systems with Values) framework, and show how political regime change promotes or demotes community values. We then present a multi-agent simulator to animate regime change with agents 'playing' the game using the AATS+V framework, using information-theoretic metrics to measure the emergence of cohesion. We conclude with a discussion of the insights that the modelling of explanations of constitutional choice provide into human behaviour and community cohesion.

I. Introduction

Cohesion is one of the most important determinants of successful and sustainable human communities and social systems, yet it appears to be one one of the hardest to define and metricate (Nowak et al., 2019). In particular, there seem to be a number of interdependent confounding factors, such as the community age, expertise and task complexity (Rychwalska et al., 2021), that affect whether a group of independent and autonomous actors, with possibly competing individual goals, can 'get together' and 'cohere', in order to address and resolve repeated collective actions problems, especially over generations.

However, Ostrom's fieldwork led to a theory of *self-governing institutions* (Ostrom, 1990), which shows how communities can self-organise a system of rules (i.e., an institution) for sustainable common-pool resource management over extended periods of time. One of the key features of Ostrom's institutions was *congruence*, the ability to adapt the institutional rules to fit the prevailing environmental conditions. Moreover, Ober's study of the classical Athenian city state (Ober, 2008) shows how Athens outlasted and outperformed its competitors, despite parity in manpower, resources, technology, etc., due to its democratic knowledge management processes that were key to successful public action. In addition, Graeber and Wengrow's challenge to the orthodox narrative of the rise of the nation-state (Graeber and Wengrow, 2022) reveals how indigenous tribes in North America were able to switch seamlessly between centralised and decentralised political regimes (with appropriate checks and balances on power) according to seasonal variation.

In the light of this work, we posit that there is some behavioural connection between the emergence of coherence and community cohesion on the one hand, and the self-determination, explanation and justification of social arrangements on the other. Since these social arrangements define the nature and style of governance or political regime operative in the community, then they affect all aspects of self-determination: (voluntary) association, constitutional choice, and their *explanation* or *justification*. In particular, constitutional variation can produce or promote (or reduce and demote) qualitative human values, such as safety, efficiency, inclusivity, participation, affinity, accountability and dignity, which can impact cohesion.

Therefore, this workshop paper lays out an investigative 'roadmap', using agent-based modelling targeted at deriving a deeper insight into this posited behavioural connection between cohesion and constitutional choice. Specifically, we propose to use the formal framework of AATS+V (Action Based Alternating Transition Systems with Values) (Atkinson and Bench-Capon, 2018) and multi-agent system (MAS) simulation to (respectively) represent and reason about self-determining actions and their effects on social arrangements. We also propose to use an information-theoretic framework to 'measure' cohesion as an emergent product of explanation and justification of decisions with respect to questions of voluntary association and constitutional choice.

Accordingly, this paper is structured as follows. We discuss the background and idea of constitutional choice in more detail in Section II. Section III uses this analysis to define the regime change game, and formalises this in the framework of AATS+V, in order to represent and reason about actions and their effects with respect to issues of constitutional choice, and how this can be used for *explainability*, e.g. for changes to be congruent with environmental circumstances or to encourage continued voluntary association. In Section IV, we consider how the AATS+V specification of the regime change game can be animated using a multi-agent simulator, and the emergence of cohesions measured using information-theoretic metrics. We conclude in Section V with a discussion of the insights that modelling explanations of constitutional choice provide into human behaviour and cohesion in communities.

II. Background: Constitutional Choice

This section gives an overview of constitituional choice, and raises some critical questions in regime change.

II.i. Constitutional Choice

In Elinor Ostrom's theory of self-governing institutions for sustainable common-pool resource management (Ostrom, 1990), social arrangements were stratified into three levels. At the lowest level were rules of operational choice, dealing with matters of resource access and distribution. These were nested within rules of collective choice, concerned with the selection, modification and application of the operational choice rules. These were in turn nested within rules of constitutional choice, i.e. rules dealing with articles of association, the initial configuration of collective choice rules, and modification of those rules.

In this respect, constitutional choice rules can be seen as being concerned primarily with the first two phases of founding in Ober's thought experiment Demopolis (Ober, 2017), namely self-selection of citizens and entrenchment of the initial arrangements. This element of self-determined social arrangements is also concerned with the nature and style of governance or political regime. They must therefore address two fundamental issues of political philosophy: first, who rules? (Plato, 1974); and secondly, how to revoke rulership? (Popper, 2002).

Some features clearly distinguish operational choice rules from constitutional choice rules. Modifying operational choice rules can be (relatively) quick, easy, have a noticeable impact which is conveniently measured by quantitative metrics, and so are more amenable to explanation or justification. By contrast, modifying constitutional choice rules is (or should be, cf. (Kahneman, 2011)) relatively slow, harder to do (e.g. through meta-rules of mutability and immutability (Suber, 1990)), appeal to qualitative values, may have conflicting drivers (Dryzek and Pickering, 2017), have an impact which may only be noticeable at a societal level or on generational timescales – and so are much less amenable to explanation and justification, in particular, in situations where a change of political regime to match environmental circumstances or a change to a societal trajectory is required.

Self-determination of constitutional choice has a particular impact on *affinity groups*, whose members are as much concerned with human relationships as they are with the problem facing the group (Bookchin, 2004). Therefore political regimes, and regime change, are a significant concern.

II.ii. Regime Change

There are many different types of political regimes, as evidenced by the variety of words with *-ocracy* or *-archy* suffixes. However, these can be categorised, as per Ober (2017), according to the answer to the question *who rules*? The answer could be broadly drawn from four options:

- one, i.e. an individual (monarchy, autocracy, etc.); or
- few, i.e. a small and exclusive coalition selected according to some specific criteria (aristocracy, oligarchy, etc.); or
- many, i.e. an extensive and inclusive body of citizens (democracy, majoritarian tyranny, etc.); or
- external authority, i.e. all decision-making is delegated to some entity outside the system, which may not be affected by the social arrangements at all.

Therefore, the first question for voluntary association to address is essentially concerned with joining or negotiating social arrangements which fall into one or other category; and secondly to address operationalisation, deciding whether or not to switch between political regimes according to prevailing conditions.

Deciding such questions may require evaluating the political regime according to some metrics concerning priorities put on personal values. Alternatively, there are *democratic indices* which purport to measure the quality of 'democratic' governance with respect to several principles, e.g. equality (good), corruption (bad), right to protest, press freedom, etc. Equally, the field of procedural justice (Rawls, 1971) tries to evaluate a political regime or institution according to how 'fairly' its procedures treat its citizens or members. Dryzek and Pickering (2017), argued that reflexive governance of ecological systems was concerned with balancing four pairs of opposing systemic drivers: flexibility vs. stability, centralisation vs. polycentricity, diversity vs. conformity, and expertise vs. engagement. A discussion of metrics for the different drivers is given in (Mertzani and Pitt, 2022).

However, at the root of any such evaluation are the different *values* that are promoted or demoted by each regime, or change between regimes; indeed political argumentation can be seen as a compromise on policies relative to different priorities or preferences on values, grounded in mutually-agreed facts or evidence. Some of the possible values which might be points of disagreement are:

- *safety*: the priority for any system of governance is the safety of its citizens (Cicero); welfare and security are two of the three fundamental provisions of Basic Democracy (Ober, 2017), the third being the avoidance of tyranny;

- *cognitive efficiency*: how much of their cognitive resources do citizens have to expend on matters of political discourse as opposed to other socially productive efforts, see for example the role of social influence in distributed information processing (Nowak et al., 2019);
- *inclusivity*: the extent, in terms of opportunity and actuality, that citizens are engaged in selecting, modifying and enforcing their chosen social arrangements, cf. Ostrom's (1990) third principle of self-governing institutions;
- *participation*: the principle that, as per (Ober, 2017), citizens should participate, and be able to participate, equally in matters of political concern;
- *accountability*: to what extent are decision-makers disproportionate beneficiaries of their decisions, to what extent are they be rewarded/punished for correct/incorrect decisions, and to what extent does accountability contribute to systemic self-improvement;
- *dignity*: although abstract civic dignity is essential; it is increased when citizens are treated as equal participants in political processes, and diminished when citizens are tricked into making decisions which they would not have made with knowledge of 'the facts' (Ober, 2017).

Although, obviously, there are specific nuances within each of the four categories of political regime – for example there is a substantive difference between a brutal tyrant and a benevolent dictator, and a democracy that fails to separate factional issues from partial good questions.

III. The Regime Change Game and AATS+V

In this section, we use an AATS+V to showcase how it can be used to explain the promotion and demotion of values in transitioning from one political regime to another when agents perform co-dependent actions. Each agent of an MAS can use an AATS+V to employ practical reasoning, and an abductive form of reasoning, to guide its own decisions that may not only depend on what others do, but also on the social consequences of making such decisions.

III.i. Regime Change Game

What we call the *Regime Change Game* is 'played' as follows. We assume that there is a starting point for a group of agents which has no form of governance. Some subset (or equals) of those agents must agree to band together in a collective (community, society, etc.), and select a preferred political from a a set of several options (as outlined above, with additional parameter settings). This implies that the agents in the collective must make a decision to perform what is known as a *joint action*.

Joint actions allow agents to move between states that represent either states either a form of non-governance $q0$, or different forms of governance. If the agents cannot agree on performing a joint action, they will not be able to transition as a society from one form of governance or non-governance to another. In such societies, we assume that there are 2 player types, namely *Proposers* (P) and *Responders* (R), where P propose a governance regime, and R accept or reject that regime.

The *Regime Change Game* can be fully specified using the AATS+V framwwork, as introduced next.

III.ii. The AATS+V Framework

Action Based Alternating Transition Systems with Values (AATS+V) are a model for rigorously reasoning about actions and their effects. The AATS+V is a well-defined structure for representing how the actions of an agent that are dependent on the actions and belief system (the values they hold) of other agents in the system will lead to transitions from one state of the multi-agent system to another.

Definition 1. (Atkinson and Bench-Capon, 2016) An Action-based Alternating Transition System with Values (AATS+V) S is defined as a $(n + 9)$ tuple:

$$S = \langle Q, q_{start}, Ag, Ac_1, ..., Ac_n, \rho, \tau, \phi, \pi, V, \delta\rangle, \text{ where :}$$

- Q is a finite, non-empty set of states;
- $q_{start} \in Q$ is the initial state;
- $Ag = 1, ..., n$ is a finite non-empty set of agents;
- Ac_i is a finite, non-empty set of actions, for each $ag_i \in Ag$ where $Ac_i \cap Ac_j = \emptyset$ for all $ag_i \neq ag_j \in Ag$;
- $\rho : Ac_{ag} \to 2^Q$ is an action pre-condition function, which for each action $\alpha \in Ac_{ag}$ defines the set of states $\rho(\alpha)$ from which α may be executed;
- $\tau : Q \times J_{Ag} \to Q$ is a partial system transition function, which defines the state $\tau(q, j)$ that would result by performing joint action j from state q. The function is partial because not all joint actions executable in all states;
- Φ is a finite, non-empty set of atomic propositions;
- $\pi : Q \to 2^\Phi$ is an interpretation function, which assigns truth values to every proposition satisfied in each state.
- V is a finite, non-empty set of values;
- $\delta : Q \times Q \times V \to \{+, -, =\}$ is a valuation function which defines the status of a value $v_u \in V$, i.e. if the value is promoted (+), demoted (-), or neutral (=), that is ascribed to the transition between two states. $\delta(q_x, q_y, v_u)$ is used to label the transition between q_x and q_y with $\{+, -, =\}$ w.r.t. value v_u.

III.iii. Regime Change Game in AATS+V

Fig. 1 draws the AATS+V for a regime change game. From this 'state machine', we can derive how changes between categories of political regime could promote or demote values, as illustrated in Tables 1 and 2.

Table 1 summarises the effect of joint actions for the respective game. It labels the regime values along with the

Table 1: Value Promotion and Demotion in the Regime Change Game Joint Actions From Non-Governance

JointAction	From	To	Proposal	Response	Promoted	Demoted
j1	$q0$	$q1$	many	accept	V1, V4, V5, V6	V2, V3
j2	$q0$	$q0$	many	reject	V4, V6	V1, V5
j3	$q0$	$q2$	few	accept	V2, V4, V5, V6	V1, V3
j4	$q0$	$q0$	few	reject	V4, V6	V1, V2
j5	$q0$	$q3$	one	accept	V1, V2, V3	V4, V5, V5
j6	$q0$	$q0$	one	reject	V4, V6	V1, V2, V3
j7	$q0$	$q4$	external	accept	V3, V4	V1, V2, V5, V6
j8	$q0$	$q0$	external	reject	V1,V2,V5,V6	V3, V4

Table 2: Value Promotion and Demotion in the Regime Change Game Joint Actions from Governance of many

JointAction	From	To	Proposal	Response	Promoted	Demoted
j1	$q1$	$q2$	few	accept	V3, V2	V1, V4, V5, V6
j2	$q1$	$q1$	few	reject		
j5	$q1$	$q3$	one	accept	V2, V3	V1, V4, V5, V6
j6	$q1$	$q1$	one	reject		
j3	$q1$	$q4$	external	accept	V3	V1, V2, V5, V6
j4	$q1$	$q1$	external	reject		

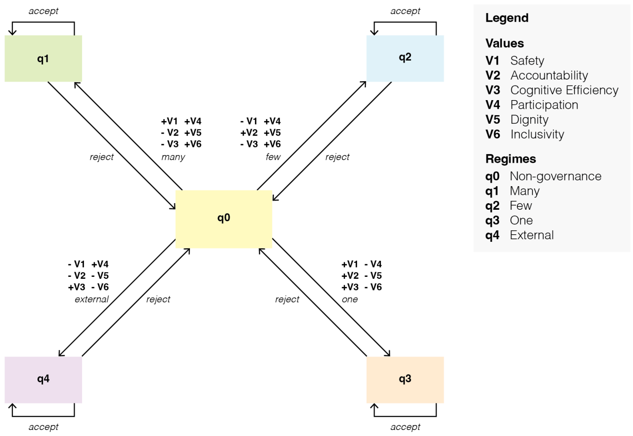

Figure 1: Transitions from $q0$, i.e., non-governance

states of the AATS+V which represent the regimes (see Legend of Fig. 1). The edges between the states represent the actions that the Proposers and Responders can perform, namely to propose a regime to change the current governance or non-governance, and to either accept or reject the proposed change.

This AATS+V considers the uncertainty of possible actions from states that represent certain political regimes. One example is q_4 in Fig. 1 from which it is unknown if agents will be able to reject a fully external regime and promote values. There is simply no AATS+V that can be drawn for deliberating in $q4$ because it is a process that takes place outside of the MAS, i.e. it is a state where the 'self-governing' property of the MAS does not hold.

Table 1 represents all the joint actions that drive the transitions and value promotion and demotion from a state of non-governance, whereas Table 2 shows us the same process from a *many* state of governance. Note that numerical metrics can be misleading: 'many' and 'external authority' ranks first equals for participation because both maximal and minimal (zero) participation is equal (if not equitable).

III.iv. Regime Change Narratives

The AATS+V enables interpretation and explanation. The AATS+V acts as a transparent operational layer on the MAS decision-making and interaction dynamics, i.e. it makes the system **interpretable**. Second, it makes the MAS simulation **explainable**, as the interpretable operational layer can generate narratives of regime changes by introducing three new elements: time T, history H, and possible histories Γ.

A history of what happened in the MAS can be extracted (see Def. 2) which can be used for justifying decisions made in the MAS by enacting the **due process** (Hewitt, 1986). Moreover, the most likely narratives that may unfold from different states in the MAS can be generated (see Def. 3), which goes beyond 'forecasting' the state where an agent population may end and justifying decisions that lead to that end state, and enables abductive reasoning – **backcasting**. This is the idea that given a preferred end-state for the MAS, we can see which decisions, events and agent-agent relations and configurations are required at particular times in order to reach that state (Dreborg, 1996).

Definition 2. A Regime Change Narrative in an AATS+V is a history, i.e., paired list of states and joint actions performed over T iterations $H_{q_n} = \{(q_a, j_b), ...(q_z, j_x)\}$, where $H \in \{Q \times J_{Ag}\}^T$.

Definition 3. Alternative Regime Change Narratives in an AATS+V is a set of histories that describe the possible transitions from a state q_x to another state q_y over a period of time T: a (3+n)-tuple $\Gamma_{q_y} = \langle q_x, T, \theta, H_i, ...H_n \rangle$, where θ is a likelihood threshold that filters the n most likely possible histories.

Clearly, having a transparent model for driving practical decision-making w.r.t. regime change is useful for justifying decisions, and explaining how to reach a desired state. Yet, a crucial element regarding the modelling of social phenomena in MAS missing from the AATS+V framework is system dynamics. A *many*-type of regime might not promote the value of *safety* forever, or the promotion of *safety* might only be expressed as an argument, while in reality having such a regime in particular contexts or in particular states of an MAS may demote safety.

This problem is not just about uncertainty, for which probabilities can be integrated to compute expected utilities of performing joint actions (Atkinson and Bench-Capon, 2016). It also has to do with some notion of stability (Ashby, 1952), since what needs to be considered are those cases where the promotion and demotion of social values are not just uncertain, but are ever-changing along with the types of agents in an MAS. It is not just a point that is being evaluated, but a trajectory (in an ever-changing field).

IV. The *Megabike* Simulator

To analyse the effect of regime change on cohesion, we propose complementing the AATS+V framework within *self-governing multi-agent system* (SGMAS). AATS+V add an explainable layer to SGMAS models, as a particular property of AATS+Vs to identify justifications. We can explain not just what happened according to the SGMAS rules, but also justify why it happened from a social perspective, e.g., the justification for switching between democratic and totalitarian regimes based on the need for rapid decision-making and the appropriate checks and balances on absolute power

IV.i. The *Megabike* Scenario

It is something of a 'tradition' in multi-agent systems to develop a simulated environment to investigate the effect of interacting cognitive agents, in order to shed light on human social interaction. The simulated environment that we are proposing to examine social cohesion through the explanation of constitutional choice is the *Megabike* Scenario, inspired by real-world multi-user bicycles.

The *MegaBike* scenario involves a group of agents which have to split into sub-groups, after which each sub-group takes control of its own *megabike*. A *megabike* is analagous to a multi-bike except it has space for up to n agents (n potentially unlimited); and each agent has its own steering wheel (unlike multi-bikes), its own pedals to propel the *megabike*, and its own brakes.

Each *megabike* must then navigate a virtual environment in search of *lootboxes* which give energy they need to survive, while avoiding an existential threat which can destroy the bike, as illustrated in Figure 2. Note that in this Figure, $n = 8$, and the colour of each agent matches the colour of lootbox from which they need energy, creating multiple conflicting incentives for free-riding, majoritarianism, cooperative survival, dynamic (re-)planning, and co-operation, coordination and competition at multiple scales.

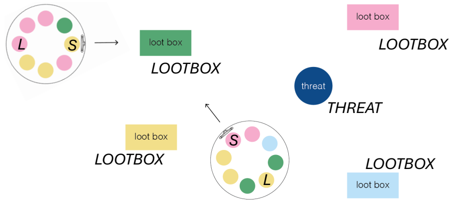

Figure 2: A visualisation of the Megabike scenario

In terms of self-govenance nd constitutional choice, the agents to follow the three stages of Ober's thought experiment Demopolis (Ober, 2017):

- self-selection: the group voluntarily divides into sub-groups, with an upper limit n on size of sub-group;
- entrenchment: each sub-group mutually agrees its own initial set of social arrangements (i.e., its constitution);
- operationalisation: each sub-group navigaets the 'world', applying and modifying these social arrangements.

Regarding the constitution, the agents have agree on rules for selecting and modifying rules, appointment to roles, membership (admission and exclusion), monitoring and sanctions, and other operational choice rules – i.e. all the functions of an institution as defined by Ostrom (1990). Note that with all the options and parameters for defining these rules, it is possible to cover the spectrum of political regimes described above, from one agent, to few, to many, to delegation to external authority.

A simulation of the *Megabike* scenario then proceeds in a sequence of iterations. Each iteration consists of two phases: in the first stage the agents apply the Demopolis self-selection and entrenchment stages. In the second phase there is the operationalisation stage, which proceeds sequentially in a series of rounds.

At the start of each round, the agents must apply their operational choice rules to collectively agree on the target lootbox, direction of travel and the speed of the bike (e.g. to avoid the existential crisis, or if they appear to be in competition with another megabike for their preferred lootbox). Then, each agent must individually decide on to the pedalling, braking or turning force it will apply. Agents have limited energy, and they expend this energy through performing these actions (brake, pedal etc.). Energy can be replenished by accessing a lootbox, which appear randomly, persist for a random number of rounds, and if un-collected disappear. Thus each agent's basic goal is to survive for as long as possible by using and restoring its energy.

Finally, the agents must review membership and reflect on the constitutional choice decisions. Effectively, the agents are faced with a regime change game: and have to hold each other accountable. They will also need to have separated out the partial goods questions, i.e. operational choice about lootbox distribution, and the factional questions, i.e. constitutional choice about rules, as per the "bright lines" principle of Democracy-by-Design (Pitt and Ober, 2018).

Note that a distinguishing feature of this scenario compared to other MAS gridworld testbeds is an interdependence of 'institutional' and 'physical' actions. For example. the constitutions might demand appointments to roles, e.g., a *leader* (L in Figure 2) who makes institutional decisions, i.e. declaring what *counts as* the target lootbox and pedalling effort, while only a *steerer* (S in Figure 2) controls its wheel for physical navigation (shortest path to target avoiding existential threat). Institutionalised power and physical capability can be represented and reasoned with using action languages Artikis et al. (2009). It also opens up identifying responsibility for 'bad' outcomes according to poor decisions or non-compliance.

IV.ii. Constitutional Choice and Emergence

Finally, we introduce an investigative framework to analyse the behavioural connection between constitutional choice (i.e. the self-determination of social arrangements) and the emergence of community cohesion. There are three parts to the framework. Firstly, agents that are capable of representing, reasoning about, participating in, and explaining the results of the regime change game. Secondly, a social network that facilitates the agent community to act as an optimised distributed information processing unit, which consistently (a) makes the 'right' decision and (b) to the agents' satisfaction. Thirdly, a way of measuring cohesion as a product of the actions and interactions pertaining to constitutional choice. In the following, we discuss each part in turn.

Agent Requirements For cohesion to emerge, agents must be able to engage in explanatory processes in order to justify the decisions they have made, they are making or the ones they are planning to make as both individuals and groups; i.e. agents must be able to **generate**, **communicate**, as well as **understand** both their own and others' explanations. These three intertwined processes imply the following requirements for the Megabike agent 'capability set':

Inference to the best explanation (IBE). IBE concerns decisions regarding rules and norms (Atkinson et al., 2020). It is an ideal form of abductive reasoning for generating hypotheses that explain some phenomenon in terms of cause and effect. The process has the following components: hypothesis generation; evidence anchoring – understanding how the observed data confirms or falsifies a hypothesis; counterfactual reasoning – changing some observed data to test or re-generate hypothesis (what if?), and contrastive reasoning – comparing the differences between hypotheses and select the fittest one.

Theory-of-Mind (ToM). ToM is concerned with the ability of agents to form and reason about models of other agents' minds (Gallese and Goldman, 1998). These models include things like what the other agents know, what they don't know, how they reason about what they know/believe, and how they perceive the environment and their peers, e.g. cognitive biases. ToM ties into explanations in two ways. First, in order to be able to generate explanations about complex dynamics and relations in the megabike, agents must be able to have some knowledge about others in a distributed manner, e.g. if another agent's preference are aligned with one's own. A simple example would be a voting justification "I voted for agent A to drive the megabike because I expect agent A to go for the green box first, which is my own preference.". Second, ToM is essential for effective communication (Miller, 2019).

Rhetoric. Rhetoric concerns effective human-human communication (Sperber and Wilson, 1990). Humans are capable of knowing what epistemic consequences their speech acts have on the beliefs of others. The role of ToM in AI agent design is to provide a model for simulating the delivery/communication of explanations as speech acts between agents that have different psychological (cognitive, epistemic, and behavioural) profiles. These ToM models should enable agents to consider who is communicating (ethos), what is being communicated (logos), how it is being communicated (pathos), and when it is communicated (kairos). Accepting or rejecting explanations depend on how all these rhetorical play into agents' fast or slow thinking (Kahneman, 2011).

One way to implement IBE capable agents is to use agent-level AATS+Vs as part of their architectures. While in Section III.iv we introduced AATS+V to explain the MAS dynamics, i.e. explaining the SGMAS for the human observer, here it would be used for agents to explain things to themselves and each other - inner MAS explanation. Considering that these agents should also use ToM to perform IBE, the AATS+V can represent higher-level human values. However, when it comes to ToM, one might require a more in-depth representation of an agent's cognitive processes. One solution would be to apply the AI architecture proposed by Lewis and Sarkadi (2024), e.g. to implement what they call the Tier 1 Reflective AI agents that use ToM to perform IBE, and then enable these to communicate with each other as demonstrated by da Silva et al. (2024).

Network Requirements Given that there are agents in the community capable of reasoning about and explaining constitutional choice, then there are two possibilities. Firstly, there are are some agents in the community that do not have this 'cognitive' competence, but can understand explanations and have representation in decision-making processes. Secondly, there are are some agents in the community that do have this 'cognitive' competence, can understand explanations and have representation, etc., but for reasons of cognitive efficiency prefer not to exercise this capacity.

Whichever of the two possibilities, collectively the agent community has to decide if the outcome of the regime change game, or its explanation, was 'right' (justified) and 'satisfactory'. This is a matter of knowledge aggregation given cognitive biases (Kahneman, 2011) and therefore social influence (Nowak et al., 2019).

The rules of social influence normally specify the principles by which a source can influence thinking and decisions of a target and overcome the passivity or resistance of the target. Implicit in this perspective is the assumption that the source's agenda is not shared by the target and is beneficial to the source. Then, social influence is closer to domination and manipulation than cohesion or 'community building'.

However, social influence may actually be beneficial to the *target*. The observation that influence may serve the interests of the target of influence underpins the Regulatory Theory of Social Influence (RTSI, Nowak et al. (2019)). RTSI is predicated on an individual's desire to be influenced and actively search for sources. Form the target's perspective, social influence is then tantamount to the delegation of information gathering and processing to others. The target chooses the topic, form and sources of influence in order to form a judgment or reach a decision on an issue. Delegating information processing to others is then functional: it saves processing resources, and can improve the quality of the decision or judgment through the emergence of expertise. It does, however, increase risk: an individual may be misled, exploited, or receive information or advice of poor quality.

The demand for efficiency pushes individuals toward delegating; but risk avoidance induces individuals to gather and process the information themselves. Therefore a critical aspect of cohesion is the balance between trust, confidence in and affinity towards expertise (Mertzani et al., 2022) and the risk of deception (Sarkadi, 2024).

Measuring the Emergence of Cohesion Information theory is an important tool in quantifying non-linear interdependencies between different components in self-organising complex systems (Rosas et al., 2019). In particular, mutual information (MI) between two variables quantifies how much knowing one variable reduces uncertainty about the other. In larger systems, though, we may want to use knowledge about a set of n *source* variables, $X_1, \ldots, X_n$, to predict a *target* variable Y. To address this need, *partial information decomposition* (PID) (Williams and Beer, 2010) proposes a non-negative decomposition of the mutual information in three types of information components or *atoms*: unique, redundant and synergistic.

In the Megabike scenario, we can study the MI between the narrative histories of n agents' *actions* as source variables $X_1, \ldots, X_n$ (e.g. repeated voluntary association, agreement over regime changes), and the temporal evolution of the emergent social cohesion as target Y, which may be quantified by the synergy between the agents' average value judgements of correctness and satisfaction (e.g. affinity between group members, or subjective assessment of treatment by rules or rulers).

Moreover, we aim to use three different principles of information theoretic causal emergence based on PID (Rosas et al., 2020), which can better reveal the relationships between the *micro* (agent) and *macro* (system) scales in the complex system being studied:

- **Causal emergence**, quantified by Ψ, refers to the property of a system that is irreducible to the sum of the system's components. This is often approximated as a synergy-redundancy index, where the emergent information is that which exists in the whole system after subtracting the information shared by the parts.
- **Downwards causation**, quantified by Δ, refers to a system feature that has a causal effect over one or more particular agents, which cannot be attributed to any other individual or group of agents.
- **Causal decoupling**, quantified by Γ, refers to a feature that can predict its own evolution, but no agent or group of agents may predict the evolution of any other element. (Note that the symbol for causal decoupling is the same as the symbol for historical narratives (Defn. 3), but they are derived from different fields and denote different objects.)

Through this interdisciplinary approach, we hope a framework of causal emergence can help identify how social values such as social cohesion causally emerge from collective actions of the agents, and whether these social values have, in turn, a feedback effect on individual action.

V. Summary and Conclusions

In summary, this paper has addressed the problem of explaining constitutional choice in self-governing multi-agent systems, in particular, explaining not just what *why* one particular configuration of voluntary social arrangements is preferable (or in some sense "better") to others, and justifying *why* it might be necessary to change from a current configuration to another.

We noted that there were numerous abstract problems with these explanations: that they appeal to abstract values (which might be hard to encode in a utility function or preference relations), and there is intrinsic uncertainty because of the extent to which change might promote some and demote other values. It also emphasises the trade-off between stability and flexibility: the ability to stay the same and the ability to change as necessary. This shows that apparent dichotomies, for example between centralisation and decentralisation, can be misleading; while even flexibility exposes a potential downside in allowing democratic backsliding.

Therefore the contributions of this paper are:

- to analyse regime change, and the values that may be promoted or demoted when an SGMAS changes from one political regime to another;
- to introduce the *regime change game*, in which an agent can use AATS+V as an abductive form of reasoning to guide its decisions about regime change preferences; and
- to define an investigative framework for implementing and animating the regime change game in an SGMAS, and demonstrating explanations of regime change.

This investigative framework opens the possibility of analysing the extent to which self-determination of social arrangements and the (quality of) explanations that agents give to each other for regime change impact on the emergence of community cohesion. Agent-based modelling of such constitutional choice could provide significant insight into human behaviour with respect to community cohesion.

Moreover, one of the critical challenges for *explainability* will be when we extend, as we plan to, the *Megabike* scenario to human-in-the-loop experiments. This will include both humans and intelligent agents interacting, joining, negotiating the social arrangements, and acting on the same *megabike*. The issue then is acceptability: can humans understand the explanations they are given, recognise the value-oriented primacy of the justifications, but most of all *accept as legitimate* decisions that might even go against them – how will they feel if they are excluded from a *megabike* by a group of agents? Will they demand that they have a human right to a human decision (Tasoulias, 2022)?

References

Artikis, A., Sergot, M., and Pitt, J. (2009). Specifying norm-governed computational societies. *ACM Trans. on Computational Logic*, 10(1):1–42.

Ashby, W. (1952). *Design for a Brain*. London: Chapman-Hall.

Atkinson, K. and Bench-Capon, T. (2016). Value based reasoning and the actions of others. In *ECAI 2016*, pages 680–688. IOS Press.

Atkinson, K. and Bench-Capon, T. (2018). Taking account of the actions of others in value-based reasoning. *Artificial Intelligence*, 254:1–20.

Atkinson, K., Bench-Capon, T., and Bollegala, D. (2020). Explanation in ai and law: Past, present and future. *Artificial Intelligence*, 289:103387.

Bookchin, M. (2004). *Post-Scarcity Anarchism*. AK Press.

da Silva, H. H., Rocha, M., Trajano, G., Morales, A. S., Sarkadi, S., and Panisson, A. R. (2024). Distributed theory of mind in multi-agent systems. In *Proc. of ICAART 2024*. SciTePress.

Dreborg, K. H. (1996). Essence of backcasting. *Futures*, 28(9):813–828.

Dryzek, J. and Pickering, J. (2017). Deliberation as a catalyst for reflexive environmental governance. *Ecological Economics*, 131:353–360.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501.

Graeber, D. and Wengrow, D. (2022). *The Dawn of Everything: A New History of Humanity*. Penguin Books.

Hewitt, C. (1986). Offices are open systems. *ACM Transactions on Information Systems (TOIS)*, 4(3):271–287.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Lewis, P. R. and Sarkadi, Ş. (2024). Reflective artificial intelligence. *Minds and Machines*, 34(2):1–30.

Mertzani, A. and Pitt, J. (2022). Metrics for reflection in distributed information processing. In *Proc. 4th Int. Workshop Agent-Based Modelling Hum. Behav.(ABMHuB)*.

Mertzani, A., Pitt, J., Nowak, A., and Michalak, T. (2022). Expertise, social influence, and knowledge aggregation in distributed information processing. *Artificial Life*, 29(1):37–65.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Nowak, A., Vallacher, R., Rychwalska, A., Roszczynska, M., Ziembowicz, K., Biesaga, M., and Kacprzyk, M. (2019). *Target in control: Social influence as distributed information processing*. Cham, CH: Springer.

Ober, J. (2008). *Democracy and knowledge: Innovation and learning in classical Athens*. Princeton University Press.

Ober, J. (2017). *Demopolis: Democracy before liberalism in theory and practice*. Cambridge, UK: Cambridge University Press.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Pitt, J. and Ober, J. (2018). Democracy by design: Basic democracy and the self-organisation of collective governance. In *IEEE SASO*, pages 20–29.

Plato (1974). *Republic*. London, UK: Penguin.

Popper, K. (2002). *The Open Society and Its Enemies: Volume 1: The Spell of Plato*. Oxford, UK: Routledge.

Rawls, J. (1971). *A Theory of Justice*. Harvard MA: Harvard University Press.

Rosas, F. E., Mediano, P. A., Gastpar, M., and Jensen, H. J. (2019). Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3):032305.

Rosas, F. E., Mediano, P. A. M., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., and Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, 16(12):e1008289.

Rychwalska, A., Roszczyńska-Kurasińska, M., Ziembowicz, K., and Pitt, J. (2021). Fitness for purpose in online communities: Community complexity framework for diagnosis and design of socio-technical systems. *Front. Psychol.*, page 12:739415.

Sarkadi, Ş. (2024). Self-governing hybrid societies and deception. *ACM Transactions on Autonomous and Adaptive Systems*, 19(2):1–24.

Sperber, D. and Wilson, D. (1990). Rhetoric and relevance. *The ends of rhetoric: History, theory, practice*, pages 140–56.

Suber, P. (1990). *The Paradox of Self-Amendment: A Study of Law, Logic, Omnipotence, and Change*. Oxford: Peter Lang Publ.

Tasoulias, J. (2022). Artificial intelligence, humanistic ethics. *Daedalus*, 151(2):232–243.

Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.