

Socially-Guided Machine Learning for Self-Organised Community Empowerment

Asimina Mertzani^a and Jeremy Pitt^b

Electrical and Electronic Engineering Dept., Imperial College London, London, U.K.
{asimina.mertzani20, j.pitt}@imperial.ac.uk

Keywords: Human-Computer Interaction, Power-Sensitive Design, Generative AI, Self-Governance, Innovation.

Abstract: Two key features of self-organising socio-technical systems are, firstly, the interaction of humans with AI, and secondly, the collective determination of social arrangements. However, this presents the risk of an inequitable distribution of power: either by translating or reinforcing existing power asymmetries directly into digital systems, unintended concession of power by human to computational, or arrogation of power by using AI as a proxy. In this paper, based on a definition of empowerment, we implement a socially-guided machine-learning system which integrates multi-agent system, generative AI and user-centred visualisation. The system is evaluated through proof-of-concept demonstrations showing how it could assist users in understanding the impact of social arrangements and so empower communities with choice, control and innovation. The significance of this work is to show how, through the synergy of human expertise, generative AI and (multi-)agent-based simulations, it might be possible to enhance human creativity to imagine original social arrangements, visualise their impact on community empowerment, and maintain an equitable distribution of power.

1 INTRODUCTION


The essence of community empowerment is that those affected by mutually-agreed, and voluntarily-complied with, *social arrangements* (rules, procedures, structures, etc.) should participate in the selection, modification and application of those arrangements. However, humans, in general, have limited experience of and expertise in the practise of determination on issues of public interest. Also, there are intricate interplays between community experience, social arrangement and task complexity (Rychwalska et al., 2021). Those in combination with the increasing hybridisation (i.e. communities involving seamless interactions between humans and AI) (Sarkadi, 2024) might produce either of two possible outcomes. The transition could be harmonious, and productive, even if in unexpected ways (Metz, 2016), or it might be harmful to individuals or damaging to the social fabric in intended or unintended ways (Robbins, 2019).


For example, the role of technology in reproducing a kind of feudalism has been observed (e.g. (Zarkadakis, 2020)). This ‘techno-feudalism’ can be attributed, in part, to an inequitable distribution of

power. This distribution may be a product of an inadvertent concession by “dumbing down” in the face of a supposedly superior intelligence (Robbins, 2022). However, it could also be a deliberate arrogation of power by using AI as a proxy for reproducing, reinforcing and amplifying extant power imbalances in socio-technical systems (Lewis et al., 2021).

This paper aims to avoid such harmful outcomes, while also providing beneficial ones, by offering a novel tool for effective self-governance through the co-production between humans and AI. This applies the *socially-guided machine learning* methodology (SGML) (Thomaz and Breazeal, 2006) to combine *codified social knowledge* and *human expertise* with *Generative AI (GenAI)* and *multi-agent simulation (MAS)* in a system for opportunistic self-organisation of innovative social arrangements for community empowerment. This derives from the use of GenAI for unexpected linkage of diverse knowledge (Metz, 2016), and MAS for unexpected emergence of pro-social behaviours (Mertzani et al., 2022).

Accordingly, this paper is structured as follows. Section 2 elaborates on the background and motivation, with respect to power and empowerment, collective deliberation, and SGML. Section 3 gives an overview of the system for community empowerment; while Section 4 details the system implemen-

^a  <https://orcid.org/0000-0002-6084-9212>

^b  <https://orcid.org/0000-0003-4312-8904>

tation. This is evaluated through proof-of-concept demonstrations in Section 5 showing how the system could assist users in understanding the impact of social arrangements and so empower communities with choice, control and innovation. After a comparative discussion of related and further work in Section 6, we conclude in Section 7 with the contributions to *interoceptive awareness* and *power-sensitive design*, through which it might be possible enhance human creativity to imagine original social arrangements, visualise their impact on community empowerment, and maintain an equitable distribution of power.

2 BACKGROUND AND MOTIVATION

2.1 Motivational Example

Open-plan offices are environments where multiple individuals share a common space, while the productivity of those individuals can be affected by the behaviour of others. Previous work has proposed the use of an anonymous online system for flagging violations of the social norms to restore human relationships and improve co-working conditions (Santos and Pitt, 2014). This was implemented in an “affective conditioning system” which combined elements of normative, affective, and adaptive computing to support self-regulation of a co-working space. Some experiments showed that psychological theories of forgiveness could be used to restore a homeostatic equilibrium after a normative violation, but was also dependent on pre-existing pro-social relationships.

With the further development of Internet of Things (IOT) and AI, the ambient environment of an open-plan office has become a socio-technical situation, in which humans and AI co-exist. This means that humans and AI interact to make decisions relevant to the self-governance of the co-working space. Such decisions affect the social arrangements (SAs) which aim to satisfy the individual preferences of the humans while considering the requirements of the technical system (e.g. minimised energy consumption). Different SAs might be agreed based on the characteristics of the individuals, the capabilities of the technology, and the nature as well as the conditions of the environment in which the space is situated.

In a simple scenario, humans might agree on a fixed set of rules concerning the behaviour of the individuals and the operation of technology. For instance, they might agree to *have their phones on silent and do not have meetings* in the shared space to decrease noise levels, and to set the *temperature equal to the*

average preferred temperature of the individuals sharing the space, and the air-conditioning (A/C) should be turned on from nine to five. Accordingly, the air-conditioning system would be set on the fixed temperature operating during the agreed times, while people would put their phones on silent by the time they get in and book a meeting room for having meetings.

However, individuals might realise that the agreed temperature should be adjusted to the seasonal variations, so they might propose a new rule, that is the change of temperature every month. It might then be observed that the air-conditioning is on while nobody is in the office. This might result in a new SA that detects motion in the space and turning on or off the A/C as appropriate. Later, they might realise that not all the people are concurrently in the space. Consequently, they might propose another SA that considers the preferences of the humans present at one time.

Overall, though, this scenario demonstrates that there is a transition underway, as physical spaces are increasingly saturated with sensors, and devices, which can exhibit some form of intelligence. Moreover, the interaction between human intelligence and this computational intelligence is dynamic rather than static, focuses on peer deliberation rather than query-answer, and involves co-production rather than provision. These features impact on issues of power, empowerment, and the self-determination of SAs.

2.2 Power and Empowerment

The previous section grounded the problem in a scenario in which humans and AI co-exist in the same environment, highlighting the need for innovating social arrangements for empowering communities to self-determine their self-organisation. The primary motivation for this work is to achieve an equitable distribution of power in contemporary socio-technical systems. However, to assess equity, we need to define the ‘measurable’ form(s) of power and empowerment.

Using modal logic, a formal characterisation of *institutionalised power* was given in (Jones and Sergot, 1996). This formalised the idea that an agent, occupying a designated role in an institution, could create facts of conventional significance by the performance of specific acts (e.g. a speech act), which “counted-as” if the institution itself had done it. The equitable distribution of institutionalised power among the agents in a MAS, was a key feature of a framework for procedural justice specified in (Pitt et al., 2013).

This characterisation of “power” is precise and even countable, but it is too narrow in the context of general SAs. However, the terms *power* and *empowerment* have been studied and defined in many

contexts, and not without contention (Adams, 2008). Summarily, there are many different aspects, including: sovereign power, interpreted as the monopoly on violence, information and charisma (Graeber and Wengrow, 2021); constitutional power, relating to the creation and framing of SAs and defining citizenship; economic power, as the accumulation of scarce resources and the leverage that provides; and even raw compute power – machine learning algorithms are computationally intensive, and democratising this technology is currently infeasible.

For the purposes of this paper, we focus on two other aspects of power. The first aspect is a subjective measurement of individual empowerment, whereby ‘intelligent’ and ‘reflective’ components in a socio-technical system have the cognitive capacity to represent and reason with respect to five cognitive dimensions of individual empowerment (referred as cognitive DoEs)(Wach et al., 2016). These dimensions are each individual’s sense of *self-determination* itself, and an *awareness* of their *competence* in, *influence* on, *knowledge* of, and *meaning* of this process, as specified in (Mertzani and Pitt, 2024).

The second aspect is an objective measurement of *collective* empowerment, referred to herein as *community health*. This is assessed by the following six properties of community health: inclusivity, transparency, diversity, equality, accountability and satisfaction (as a form of interactional justice, i.e. how well individuals feel they have been treated).

2.3 Empowerment of Deliberation

“Social arrangements”, informally introduced in (Graeber and Wengrow, 2021), is an umbrella term for any type of socially-constructed rule-based system mutually agreed by members of a group, to voluntarily regulate their behaviour and hold themselves accountable to one another. In this paper, we identify SAs as the structures, rules and processes used in a community to self-organise its empowerment, including its polity (i.e. relationships to external actors).

For example, an SA focused on improving the cognitive dimension of *knowledge* could be “Organise a monthly discussion panel related to democratic decision-making”. In a human community the extent of their knowledge could be determined (e.g. by survey) before and after introducing this SA, although it might also have an effect on other dimensions.

However, this raises two questions: firstly, from where do new SAs originate; and secondly, how can the effect of an SA be evaluated? Ideally, the answer to the first question would be endogenous: it would come from within the community itself. In

practice, given people’s limited experience and expertise in such matters, some guidance is likely to prove necessary. We propose that this task is very well-suited to GenAI, in part because the models would have been trained on large volumes of data from a variety of sources, but also, as previously mentioned, because of GenAI’s capacity to find unexpected linkages in data, which can inspire human creativity and innovation. Moreover, although the use of GenAI is of contention, it has been shown to support human creativity and brainstorming (Bouschery et al., 2023; Memmert and Tavanapour, 2023).

The answer to the second question lies in modelling and simulation. With a good model of the community, we could better understand the effect of the SA. For this purpose, MAS are an appropriate tool: we can design and implement a MAS that, to the extent that it reliably models the community, can be animated and used to evaluate the effect of an SA on the community through local interactions and social influence. As previously mentioned, a key feature is the emergence of unexpected pro-social behaviours.

2.4 Socially-Guided Machine Learning

In a variation of the model-view-controller pattern, we use a MAS model, a visualisation of the empowerment of the agents, and *two* controllers: the human user and GenAI, as illustrated in Figure 1. Either the human user or the GenAI can recommend alternative SAs: the effect of these new SAs is simulated in the MAS, and the impact on community empowerment is visualised for human ‘consumption’. The co-production of SAs between aims to confine the weaknesses of GenAI (e.g. biases, hallucinations), while benefit from its strengths to support human creativity.

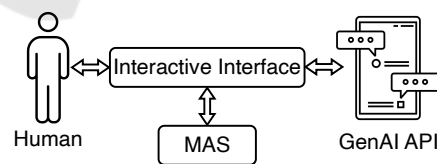


Figure 1: Socially-Guided Machine Learning.

Hence, using Socially-Guided Machine Learning (SGML) methodology (Thomaz and Breazeal, 2006), we iterate first through a phase in which we run the MAS and visualise its final state to the user, and second a phase in which the user evaluates that state and proposes a change (with or without consulting GenAI). This change is applied to the system and leads to the next iteration.

As a result, because this is essentially a non-deterministic cybernetic system whose outputs are its own inputs, what happens to the community is more

significant in determining its final state than the starting conditions. Resetting (or bootstrapping) the system allows exploration of multiple different iterations of proposed SAs, enabling evaluation of comparative performance and long-term impact with different combinations of GenAI and MAS behaviour, as well as the opportunity for potentially unbounded co-production of new SAs.

3 SYSTEM OVERVIEW

3.1 System Interface

The interface of the system aims to facilitate the interaction between the user, GenAI, and MAS. Specifically, it supports the following user activities:

- input of the SAs they want to test in the MAS
- get inspiration of new SAs by querying the GenAI
- access the results of applying an SA to the MAS through visualisations (e.g. graphs)

Specifically, an indicative example of visualisation, enabling the users access the effects of the iterative application of an SA to the MAS is shown in Figure 2. These spider plots show the extend in which each cognitive DoE and community health property is fulfilled, ranging from 0% to 100%. This interface is suitable for being accessed via a web browser, while running in a remote server, according to a client-server model, however the current version of it is hosted locally.

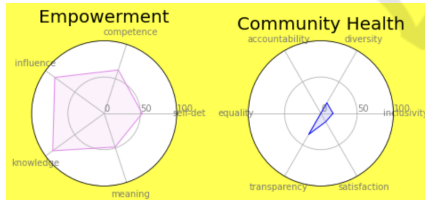


Figure 2: Example Visualisation.

3.2 Operational Cycle

The control flow of the system is described in Algorithm 1 and in Figure 4. The modules specified are analysed in the next Section. The text in green corresponds to user input, text in blue is a displayed output, and text in red is a module call. The operational cycle starts with initialisation of the MAS, initialisation of the interaction with the GenAI API, and an initial simulation run of the MAS for a predefined number of rounds m , for which no SA is proposed or used. After the completion of the first epoch, dimensions of the empowerment and community health properties are displayed by the visualiser.

Algorithm 1: Control Flow.

```

1: Let  $t = 0, epochs = 0, rounds = 0, reset = false$ 
2: System Initialiser. (Section 4.2)
3: for  $rounds = m$  do
4:   Baseline Scenario Executor. (Section 4.4)
5:   Parameter Collector.
6:    $rounds = rounds + 1, t = t + 1.$ 
7: end for
8: Visualisation  $\rightarrow$  Visualiser (Section 4.11)
9:  $rounds = 0, epochs = epochs + 1$ 
10: while  $t < T$  and not( $reset$ ) do
11:   Input1  $\rightarrow$  Ask if need for change of SA.
12:   if  $Input1 \neq No$  then
13:     Need Detector. (Section 4.5)
14:     Output1  $\rightarrow$  Need from Need Detector.
15:     if  $Input1 == Human$  then
16:       Input2  $\rightarrow$  Input of SA from User.
17:     else if  $Input1 == GenAI$  then
18:       GenAI Messenger. (Section 4.6)
19:       Output2  $\rightarrow$  SA from GenAI Messenger.
20:       Input2  $\rightarrow$  Ask for validation or modification.
21:     end if
22:     SA Analyser. (Section 4.7)
23:      $t_{sa} = t.$ 
24:     end if
25:     for  $rounds = m$  do
26:       Baseline Scenario Executor. (Section 4.4)
27:       SA Effect Calculator. (Section 4.8)
28:       DoE Calculator. (Section 4.9)
29:       Opinion Formation Executor. (Section 4.10)
30:       Parameter Collector.
31:        $t = t + 1, rounds = rounds + 1.$ 
32:     end for
33:     Output3  $\rightarrow$  Informative Message from SA Effect Calculator.
34:     Visualisation  $\rightarrow$  Visualiser (Section 4.11)
35:      $epochs = epochs + 1, rounds = 0.$ 
36:     Input3  $\rightarrow$   $reset = User$  decides to bootstrap.
37:   end while
    
```

The *Program* queries the user if s/he wants to change the SA, and replying in the affirmative, s/he has also to decide if the new SA should be user- or GenAI-generated. The system determines and displays what parameter is under-valued, and either asks the user to input a new SA or queries the GenAI; in the latter case, the user can accept or modify GenAI's answer. In either case, the outcome is a new SA; if the user did not want to change the SA, the 'new' SA in the next epoch is the same as the old SA.

The new SA is applied in another epoch of the MAS, and its effects after each round are stored in memory. After the completion of the epoch (m rounds of the MAS), the impact of the SA in each cognitive DoE and community health property are presented to the user by the visualiser.

Finally, the user is asked to consider bootstrapping the system, in which case the next cycle will start from the initialisation stage, or if not, to proceed to the next epoch, which then returns control to the original change query. Resetting the system to its original starting conditions, and re-running, enables the user, and the system, to learn what SA, and what order of SAs, have what impact in which situations.

3.3 Illustrative Walkthrough

Figure 3 provides an illustrative walkthrough of the system operation: on the left, the visualisation of empowerment and community health; and on the right, the user-system dialogue.

In the first ‘row’, the user chooses to input his/her own SA; the system recommends an SA to improve the *knowledge* DoE, and the user inputs a new SA. In the second ‘row’, the user chooses to query GenAI; the system recommends an SA to improve the *influence* DoE, and queries GenAI accordingly. The answer produced can be accepted or modified (in this case it is accepted). In the third row, the user decides to stick with the current SA. In each case, the impact of the SA on the MAS is computed and new level of empowerment, in terms of cognitive DoEs and community health properties, is demonstrated by the change in shape of the spider plots.

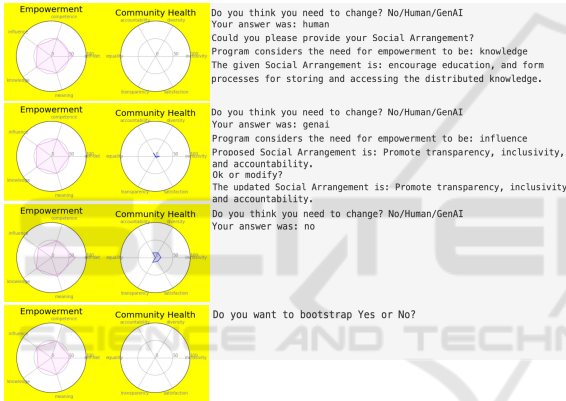


Figure 3: Illustrative Visualisation and Dialogue.

4 SYSTEM IMPLEMENTATION

This section summarises the system’s implementation, giving a detailed specification of each module, and describing their information processing.

4.1 System Architecture

The *Program* is composed of ten modules:

- the *System Initialiser*, which instatiates the MAS and the independent parameters.
- the *MAS Simulator*, which generates the MAS;
- the *Baseline Scenario Executor*, which runs the MAS for an epoch (m rounds) without any SA;
- the *Need Detector*, which evaluates the state of the MAS and determines the need for empowerment;
- the *GenAI Messenger*, which sends and receives prompts by GenAI API;

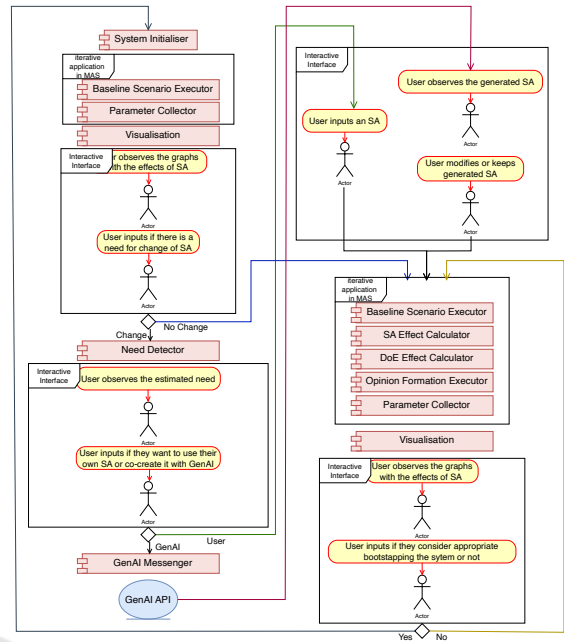


Figure 4: System’s Architecture.

- the *SA Analyser*, which extracts the information from the GenAI response or the input from the user, and maps it to DoEs;
- the *SA Effect Calculator*, which grounds an SA to the current round;
- the *DoE Calculator*, which calculates the cognitive DoEs and community health properties for the current round;
- the *Opinion Formation Executor*, in which agents interact and compute the updated cognitive DoEs and community health properties; and
- the *Visualiser*, which shows the empowerment after the iterative application of the new SA.

The way the modules process information is as follows. *System Initialiser* assigns values to the parameters according to the inputs of the user. *MAS simulator* generates the MAS according to the inputs from the Initialiser and outputs an instance of itself (e.g. a MAS comprising x agents etc.). *Baseline Scenario Executor* receives the MAS instance and calculates agents’ cognitive DoEs for m rounds without using any SA. For instance, for each agent a in round r , it calculates each cognitive DoE i , which would correspond to $cd_{a,i,base}^r = y$ (*base* is used to differentiate the baseline value of the DoE from the one after adding the contribution from the SA) aggregates them, and forms the corresponding $CD_{i,base}^r = Y$.

Need Detector receives $CD_{i,base}^r$ and computes the need, e.g. $need^r = influence$. This together with the known SAs are the inputs of the *GenAI Messenger*,

which interacts with GenAI API and outputs a new SA, e.g. ‘Foster a culture of open communication’. The new SA is sent to the *SA Analyser*, which outputs a *bonus* vector corresponding to the mapping of SAs to DoEs, e.g. $bonus = [self-determination:1, competence:0, influence:2, etc.]$. *SA Effect Calculator* initialises the distribution of the effectiveness of an SA over time and outputs the current effectiveness, for instance effectiveness of SA i in round r is ef_i^r .

Furthermore, the outputs of the *Baseline Scenario Executor*, the *SA Analyser*, and the *SA Effect Calculator* become the inputs to the *DoE Calculator* which calculates the agents’ current DoEs. For instance, for the given scenario, its output would correspond to $cd_{i,a}^r = Z$ and $CH_j^r = W$. The agents’ cognitive DoEs are sent to the *Opinion Formation Calculator* who lets agents’ interact and choose to use their $cd_{i,a}^r$ or neighbours $cd_{i,a,nei}^r$ cognitive DoEs and form their final $cd_{i,a,f}^r$. As such, the output of that module becomes the aggregate value of the agents’ cognitive DoEs $CD_{i,a}^r$ and together with the CH_j^r are sent to the *Visualiser* which outputs a graph similar to that of Figure 2. For brevity, when we refer to the DoEs of all the agents we omit a and i . Overall, the modules are summarised in Table 1, while the system’s architecture is given by Figure 4.

Table 1: Modules and Experimental Parameters.

| Modules | | |
|----------------------------|-------------------------------------|-------------------------------------|
| Name | Input | Output |
| System Initialiser | independent parameters | Instantiated independent parameters |
| MAS Simulator | Instantiated independent parameters | MAS instance |
| Baseline Scenario Executor | MAS instance | $cd_{a,base}^r, CD_{base}^r$ |
| Need Detector | CD_{base}^r | $need^t$ |
| GenAI Messenger | $need^t, KB$ | new_SA |
| SA Analyser | new_SA | $bonus$ |
| SA Effect Calculator | Gamma params., t_{sa}, t | ef_i^r |
| DoE Calculator | $cd_{i,a}^r, bonus, ef_i^r$ | cd_i, CH^r |
| Opinion Formation Executor | $cd_{i,a}^r, SN, cr, sc$ | CD^r |
| Visualiser | CD^r, CH^r | Graphs (e.g. Fig2) |

| Independent and Exp. Determined parameters | | | |
|--|--|--------|----------|
| Symbol | Description | In/ExD | Value |
| \mathcal{N} | number of agents | In | 100 |
| T, m | maximum duration in epochs and rounds in an epoch | In | 50,25 |
| KB_{init} | initial knowledge base of SAs | In | 0 |
| $cgmax, dev$ | max fixed agents’ cognitive DoE and allowed deviation | ExD | 10,- |
| $-d, d$ | agents’ cognitive DoE allowed deviation interval | ExD | [-1,1] |
| k, θ, loc | shape, scale and time shift of Gamma distr. | ExD | 5, 10, - |
| loc_f, loc_r | time shift fixed and random part | ExD | 1,- |
| loc_{min}, loc_{max} | lower/upper bounds of time shift random part | ExD | [-5,5] |
| mul | PDF multiplier of Gamma | ExD | - |
| mul_f, mul_r | mult. fixed and random parts | ExD | 50,- |
| mul_{min}, mul_{max} | lower/upper bounds of mult. random part | ExD | [-25,25] |
| $bonus_{max}$ | max. reinforcement to a DoE | ExD | 10 |
| τ | critical health message threshold | ExD | 45 |
| c | reinforcement of credence/confidence in a round | ExD | 0.01 |
| $cr_{a,n,init}, sc_{a,init}$ | initial credence of agent a in n /self-confidence of a | ExD | 0.5,0.5 |

4.2 System Initialiser

The initialisation of the system includes the instantiation of the MAS and the definition of the experimental parameters. Table 1 gives an overview of the independent and experimentally determined parameters and presents the values assigned to them in the experiments below.

4.3 MAS Definition

The MAS corresponds to a self-organising institution comprising N agents, having a knowledge base KB , a list with the five cognitive DoEs CD (where CD_i^t is the collective value of them in time t), and one having the collective DoE (corresponding to community health properties). Each agent a is initialised with a fixed individual value for each cognitive DoE cd_a (randomly elected from the interval $[0, cgmax]$, where $cgmax$ is an experimentally determined parameter) and in each round this value can deviate from the fixed by a random value dev from the interval $[-d, d], d > 0$, e.g. $cd_a^r = cd_a + dev$. This is to generate some agents being ‘experts’, i.e. having higher cognitive DoEs and consequently being more empowered. Also, each agent a has a social network, and has credence $cr_{a,n}$ to each of their neighbours n , but also self-confidence sc_a , and agents are initialised to have equal self-confidence and credence to all neighbours. Also, Table 2 provides an example of KB .

4.4 Baseline Scenario Executor

In the baseline, no SAs known, so the agents iteratively form their cd_{base} and the total value of each DoE i is given by Equation 1:

$$CD_{i,base}^r = \frac{\sum_{a \in \mathcal{N}} cd_{a,i}^r}{cgmax * \mathcal{N}} * 100 \quad (1)$$

where \mathcal{N} is the number of agents, and $cd_{a,i}^r$ is agent’s a value of each cognitive DoE i in round r . However, the collective DoEs are zero in the baseline, i.e. $CH_{base} = 0$. This is to reflect the lack of community health when there is no knowledge about SAs.

4.5 Need Detector

Need Detector senses the MAS and detects the current need in terms of SAs. Therefore, it constitutes an internal mechanism of interoceptive awareness (Pitt and Nowak, 2014) developed in the MAS, which compares the cognitive DoEs and highlights a ‘threat to the body politic’. Specifically, it receives the past cognitive DoEs, it calculates the rate of change of each $CD_{i,rate}$, and outputs the $need^t$ which corresponds to the cognitive DoE i that has the greatest negative rate of change in the current epoch t , given by Equation 2:

$$need^e = \operatorname{argmax}_i \frac{\sum_{e=1}^t (CD_i^e - CD_i^{e-1})}{t} \quad (2)$$

4.6 GenAI Messenger

The *GenAI Messenger* receives the current need $need^t$ and the KB , and composes a message in which it spec-

ifies the need and the known SAs for that need (i.e. the corresponding row in the *KB*-Table). For instance, if the need is ‘self-determination’ and the *KB* is that of Table 2 an indicative interaction between the system and the API is:

Message: “The population needs a social arrangement to improve the individuals’ cognitive dimension of self-determination, and already knows the following: ‘Promote active engagement’. Can you give another one?”

Response: “Empower individuals to have a voice and take ownership in shaping their communities.”

The message is sent to the GPT-4o mini API (OpenAI, 2024) and the reply corresponds to the *new_SA*, which is shown to the user for modification or approval.

4.7 SA Analyser

The *SA Analyser* receives the input from the user or the (user-modified or not) response from GenAI, either of which corresponds to a new SA, and outputs a vector, named *bonus*, with the expected impact of a new SA to the eleven DoEs. To convert the SA to a vector, we define a vocabulary comprising terms (e.g. words and phrases) that are mapped to DoEs. So, each time a new SA is given, it is analysed, as described below, and the terms in it are used to define the reinforcement, which is then specified in the *bonus* vector.

Table 2: Example KB and Part of Vocabulary.

| Cognitive DoE | | KB | | | | | | | | | | |
|----------------------------------|-----------------------------------|--|------|---|---|------|---|---|----|---|---|---|
| | | SAs | | | | | | | | | | |
| Self-determination Competence | Influence Knowledge Meaning | Promote active engagement. Empowerment through education and support. Promoting self-advocacy and decision-making skills. Foster a culture of open communication. | | | | | | | | | | |
| | | Encourage civic education and run several assessments Conduct surveys to gather input and address any discrepancies. | | | | | | | | | | |
| Part of Vocabulary | | | | | | | | | | | | |
| Term | S-D | C | Infl | K | M | Incl | D | A | Eq | T | | |
| Active Participation | Y | N | N | N | N | N | N | N | N | N | N | N |
| Open Communication | N | N | Y | N | N | Y | N | N | N | N | Y | N |
| Ownership | Y | N | N | N | Y | N | Y | Y | N | N | N | N |

An indicative vocabulary is given by Table 2 and the full vocabulary is available [here](#). The first column presents the terms, and the other columns correspond to the eleven DoEs, where the letter ‘Y’ denotes that the term affects the DoE while ‘N’ shows that it does not. The steps are the following:

1. the *bonus* vector assigns a zero to each DoE
2. the punctuation marks of the new SA are removed and the text is converted to lowercase
3. the SA is parsed and the analyser is looking for key-terms that are included in the vocabulary

4. if a keyword is detected, then the analyser looks for ‘Y’ in the vocabulary and increases the value of that DoE in the *bonus*.

For instance, if the *new_SA* suggests to ‘Foster a culture of open communication, collaboration, critical thinking, and ethical behaviour to engage individuals in meaningful contributions to collective knowledge, and decision-making’, then the *bonus* would be:

bonus = [self-determination: 3, competence: 2, influence: 3, knowledge: 3, meaning: 4, inclusivity: 3, transparency: 2, diversity: 0, equality: 1, accountability: 1]

4.8 SA Effect Calculator

By analogy to a community in which humans need time to process and engage with changes, the system is designed so that each time a new SA is applied in the MAS, the population requires some system time to understand it, engage with it and derive the benefits, as discussed in (Kahneman, 2011). Also, after some time, the population or the environment might change, and that SA might not be relevant anymore. So, we set the effectiveness of an SA to follow a Gamma distribution, as this is also used to model the waiting time for a drug to reach its maximum effect in the body. Moreover, since different SAs require a different amount of time to be accepted, and others might be more or less effective, we generate a new Gamma distribution for each new SA, with the following properties.

The shape k and the scale θ of the Gamma distribution of each SA are experimentally determined parameters. Moreover, the distribution is shifted on time by loc , which is a parameter having a fixed loc_f and a random loc_r . D which is sampled from a uniform distribution defined in the $[loc_{min}, loc_{max}]$ interval, and is equal to $loc = loc_f + loc_r$, where $loc_r \sim Uniform(loc_{min}, loc_{max})$, and loc_f, loc_{min} and loc_{max} are experimental parameters. Also, to make the probability density function (PDF) of the Gamma distribution to take values from minus one to one, in the specified interval, its value is multiplied by mul which has a fixed mul_f and a random mul_r . D defined similarly with the DoEs of the shift.

As such, this module takes as input the experimental parameters k , θ , loc , loc_{min} , loc_{max} , mul , mul_{min} and mul_{max} , the current t and the number of rounds since the SA application t_{sa} , and calculates the effectiveness $ef_i^{t-t_{sa}}$ of SA i in time t . This is equal to the value of the probability density function (PDF) of Gamma at the current time t minus the time that the new SA was first applied t_{sa} , multiplied by mul , and shifted by loc , given by Equation 3:

$$ef_i^{t-t_{sa}} = mul * f(t - t_{sa}; k, \theta) + loc \quad (3)$$

This module has an additional control developing a second layer of interoceptive awareness (in the human computer interaction), on top of the one developed in the *Need Detector* (in the MAS). This calculates the gradient of the distribution of the effectiveness on the current time instance after the application of the SA ($t - t_{sa}$) and makes the user aware of whether it has reached its maximum effectiveness (negative gradient) or not (positive gradient).

So, depending on the sign of $\nabla ef_i^{t-t_{sa}}$ the messages ‘*Maximum effectiveness has been reached*’ and ‘*Wait more for maximum effectiveness*’ are shown to the user. Furthermore, if the SA has reached maximum effectiveness and any of the DoEs is lower than a threshold τ , a warning message saying ‘*Community is disempowered, consider a change!*’ is shown.

4.9 DoE Calculator

With input the baseline cognitive DoEs, the *bonus* for the SA i , and its effectiveness ef_i^t , the *DoE Calculator* outputs the agents’ DoEs. The value of an agent’s a cognitive DoEs are $cd_{a,i}^t$, the populations’ cognitive DoE CD_i^t equals the average of $cd_{a,i}^t, \forall a \in \mathcal{A}$, and the collective DoEs are CH_i^t , given by:

$$cd_{a,i}^t = \frac{cd_{a,i,base}^t * bonus * ef_i^t}{bonus_{max}}, CH_i^t = \frac{bonus * ef_i^t}{bonus_{max}}$$

where *bonus* is the *bonus* of the current SA and $bonus_{max}$ is an experimental parameter that gives the maximum bonus that an SA can cause to a DoE.

4.10 Opinion Formation Executor

The *Opinion Formation Executor* is responsible for the interaction of the agents and the formation of their final opinion in terms of cognitive DoEs. In particular, agents interact with their social network and they can choose to use their own cognitive DoEs or to ask a neighbour. This is to make MAS simulate a community, in which agents interact and can choose to use their own opinion or ask a source to optimise their decision making, as described in (Nowak et al., 2019).

As such, an agent a has their cd_a^t and selects one of the most credited neighbours n for their cd_n^t (corresponding to $cd_{a,nei}^t$). Based on $cr_{a,n}$ and sc_a , a uses the own or asked cognitive DoEs in their final $cd_{a,f}$, given by Equation 4, while the population’s final cognitive DoEs CD correspond to their mean:

$$cd_{a,f} = \begin{cases} cd_{a,nei}^t, & \text{if } cr_{a,n} \geq sc_a \\ cd_a^t, & \text{if } cr_{a,n} < sc_a \end{cases} \quad (4)$$

Then, a updates the credence to n and self-confidence based on whether the average value of n ’s cognitive

DoEs is greater than a ’s. The amount of the reinforcement of those per round is defined by an experimentally determined parameter called c , and the credence reinforcement is given by Equation 5, while the opposite applies for the self-confidence:

$$cr_{a,n} = \begin{cases} cr_{a,n} * (1 + c), & \text{if } \sum_{i \in CD} cd_{i,a,nei} \geq \sum_{i \in CD} cd_{i,a} \\ cr_{a,n} * (1 - c), & \text{otherwise} \end{cases} \quad (5)$$

Note that the opinion formation affects only the agents’ cognitive DoEs and not the community health properties. This is because agents individually can optimise their cognitive capacity, but the community health is an emergent property that is the outcome of cognitive empowerment and well-being. Therefore, it cannot be optimised by simply asking a source.

4.11 Visualiser

After running the MAS for m rounds, the system has to inform the user regarding the effects of an SA to it. Therefore, the *Visualiser* receives the DoEs and generates some informative graphs. An indicative examples is that given by Figure 2.

5 EVALUATION AS PROOF OF CONCEPT

The system developed is non-deterministic, implying that the sequence of events is more significant than the starting conditions, so different order of inputs, can result in different outputs. However, this section gives proof of concept examples which demonstrate the development of interoceptive awareness, the variety of SA effects, and the role of the user and the emergence of expertise through bootstrapping.

The MAS defined for that experimentation has the minimum amount of characteristics to showcase that it can be relevant to multiple different applications. Specifically, it comprises N agents initialised with randomised values of DoEs forming a Klem-Eguiluz social network, which resembles real-life networks (Prettejohn et al., 2011), and the values of the other related parameters are given by Table 1. A more elaborate definition of the MAS in a specific context would increase the leverage of the users.

5.1 Interoceptive Awareness

To facilitate the empowerment of communities through the self-determination of SAs, the system is designed to have mechanisms that inform the users

and enable them develop interoceptive awareness with respect to the system’s health (second level interoceptive awareness mechanism).

Figure 5 showcases this property of the system, where the user initially identifies the need for change and uses GenAI (first grey box). The SA proposed is ‘establish clear communication channels, utilize collaborative tools, encourage knowledge sharing and collaboration, and regularly review and update knowledge repositories’, and is approved. This progressively results in increased empowerment, mainly in terms of knowledge, influence, diversity, inclusivity, transparency, and satisfaction. However, after some epochs, the SA is not effective anymore resulting in decreasing the DoEs, shown in the spider plots. This is observed by the system which displays a warning (highlighted in red in the Figure).

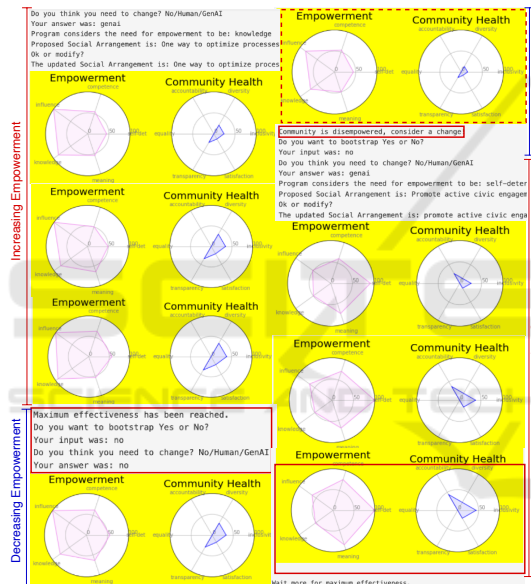


Figure 5: User’s Interoceptive Awareness.

Accordingly, the user proposes a change using GenAI, and the new SA is ‘Promote active civic engagement, and civic participation, and provide opportunities for meaningful and equal participation in decision-making.’ which is approved and results in increasing DoEs. Therefore, the system owns a mechanism of self-healing which is achieved through the effective user-computer interaction. Specifically, the system makes the user aware of the need, the user acknowledges that and triggers a change, and that change is applied to the system and enable its recovery from a state of disempowerment.

5.2 Variety of SA Effects

As discussed above, different SAs result in different outcomes in terms of DoEs. The upper part of Figure 6 shows the DoEs in the baseline compared to those when applying two different SAs, which are ‘Establish clear communication channels, utilize collaborative tools, and regularly update and share information among team members.’ during round 600, and ‘Conduct regular surveys or feedback sessions to gather input and address any discrepancies’ during 800.

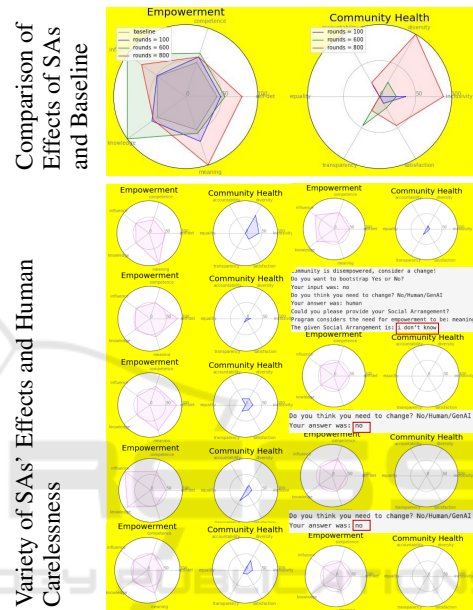


Figure 6: Variety of SAs and Human Carelessness.

Additionally, the lower part of Figure 6 shows the effect of the following SAs:

- Regularly review practices against core values and encourage open communication for dissent.
- Promote inclusive and transparent decision-making to empower individuals to actively participate in shaping their communities.
- Foster a culture of inclusivity, transparency, critical thinking, and accountability.
- Establish clear communication channels, utilize collaborative tools, and encourage knowledge sharing.

Notice that different SAs address different needs and therefore result in new DoEs, while all outperform the baseline. For example, the first SA increases mainly meaning, diversity and inclusivity, while the fourth increases influence, knowledge, and transparency. This highlights how proposed SAs should be based on the need of the system, which is also the reason why the system detects the need and uses that to optimise its decision-making with respect to SAs.

5.3 Importance of Human Commitment

Focusing on the role of the user, the right side of the second row of Figure 6 provides an example of interaction in which the user is indifferent, and therefore cannot benefit from the system’s suggestions. In particular, despite the user opting to make a change, their input is irrelevant (e.g., “I do not know”). In the next epochs, although the DoEs are decreased the user does not intervene, e.g. they do not trigger a change as highlighted in red boxes in the Figure (e.g. “no”). Overall, the carelessness of the user and their indifference to the warnings results in individual and collective disempowerment (reflected by low DoEs).

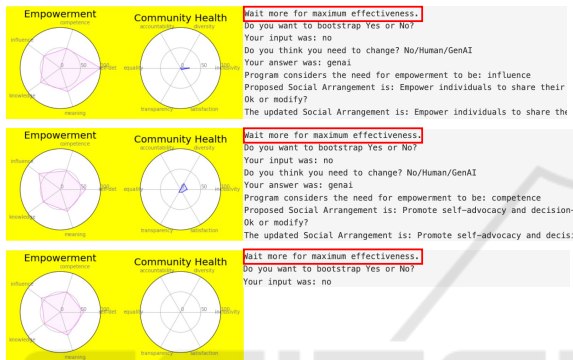


Figure 7: Human Impatience.

Another aspect for the effective operation of the system is that the user is patient. Figure 7 shows a scenario in which the user consecutively asks GenAI to propose a new SA, and they do not wait for this to be effective. Specifically, the AI generated SAs are the following:

- Encourage individuals to engage in shaping their communities through inclusive decision-making processes;
- Support people to share perspectives, listen actively, seek evidence-based solutions, and uphold ethical principles;
- Promote self-advocacy and decision-making.

Although the SAs can be beneficial for the system, its impact does not become apparent due to the user’s impatience (reflected by quick changes from SA to SA). This way the population in MAS does not manage to engage with the new SA, remaining disempowered. Therefore, it is critical that the user follows the suggestions of the system (i.e. message ‘Wait more for maximum effectiveness’) and waits till the effects of the change become measurable.

5.4 Opinion Formation in MAS

As discussed in the *Opinion Formation Executor*, agents can decide to ask a source (a neighbouring agent from the social network having higher cognitive DoEs). Therefore, this highlights the emergence of expertise and specifically shows how agents learn to distinguish the ‘experts’ and seek for their opinion (corresponding to cognitive DoEs).

This emergent property can be observed in Figure 8, where the first column shows the amount of the agents using their own (red) and their neighbours (blue) cognitive DoEs to form their final cognitive DoE, and the second row shows the population’s cognitive DoEs where they are using their own (red) compared to the ones that they finally select (blue). Therefore, the first graph shows that agents in the MAS learn to use their credited sources’ cognitive DoEs, fact that results in higher cognitive DoEs reflected by the graph in the second row. As such, it becomes evident that not only they ask their neighbours, but also identify the experts, and overall optimise their community’ cognitive capacity leading to empowerment.

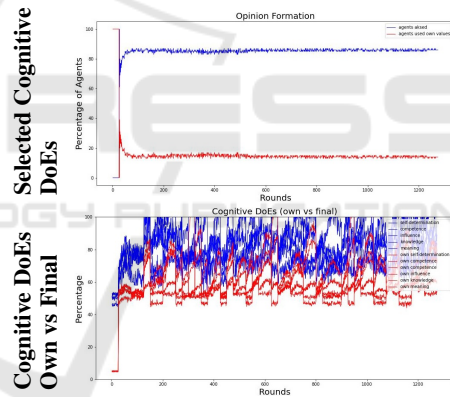


Figure 8: Effect of Social Influence in Cognitive DoEs.

6 RELATED AND FURTHER WORK

6.1 Related Work

Our work contributes to initiatives for supporting localised self-governance, evidence-based law-making and enactment, and effective human-AI co-production. This section discusses how it builds on those attempts and combines them to offer a mechanism that combines MAS, GenAI, and human users to support community empowerment.

In the field of political science, (Manville and Ober, 2019) have called for a new “release” of democ-

racy, Democracy 4.0. They propose new architectures of engagement, which includes not just more direct participation and less “representative” approaches to decision-making, but also channelling democratic energy and initiative into smaller scale, but more personally meaningful, forms of self-governance (cf. (Bookchin, 2004)). This system complements their call by providing a technology that could support this localised self-governance by making accessible cause and effect relations to the users.

In the domain of legal drafting, it is observed that “traditional drafting methodologies produce laws that do not work” (Seidman and Seidman, 2009). This failure is attributed, in part, to the focus of legislative scholarship, in the study of power, being on the process of compromise between legislators on detailed provisions before enactment. The four-step ILTAM methodology is intended to assist law-makers enact evidence-based legislation that works as intended. With our system, we are not only transferring power from legislators to those affected by legislation, but also providing evidence (using MAS) of the impact of new SAs on the extent of that empowerment.

There is awareness of the need for a better understanding of human-AI co-production (Thomaz and Breazeal, 2006), and hybrid systems integrating machine learning and machine reasoning (Kierner et al., 2023). Our work contributes to both initiatives, in the former by allowing the human and AI to work in tandem on the selection and recommendation of new SAs, and in the latter by enabling GenAI and MAS to work in tandem on the application and effect of SAs. This system also admits a potential for brainstorming and enrichment of human knowledge through repeated exploration of different event sequences being applied to the same starting conditions.

6.2 Further Work

Although substantive results have been demonstrated as a proof of concept, this system is at present a work in progress, and there is a number of limitations to overcome and improvements to be made in further research and development. These include:

- in the *SA Analyser*, we need a substantive user survey to establish a stronger correlation between keywords and the DoEs;
- in the *MAS*, a method for mapping the features and characteristics of a community to the MAS is required;
- in the *Need Detector*, we need to develop a more complex mechanism for interoceptive awareness that takes into consideration multiple parameters;

- in the *Visualiser*, although some text entry (of SA) is required, we should upgrade the text-based dialogue to improve the UX;
- generally, considering groups of users or experts could improve the system’s effectiveness, while further exploration on issues of model misalignment, scale, and change on human behaviour due to the interaction with the system is needed; and
- overall, we aim to package the system as a plug-in for PlatformOcean (Pitt et al., 2021) and evaluate performance in field trials.

Following these developments, we would aim to deploy and evaluate the system in a field trial, with several possible applications, for example in deliberative assemblies for public policy, or co-housing for local communal policy formation.

7 CONCLUSIONS

This paper has identified an inequitable distribution of power as a fundamental challenge for the development of socio-technical systems. Accordingly, it has developed a system that helps people design their communities better, using a model of equitable society as a guide. This system, enables users to visualise the extent of community empowerment in cognitive and collective dimensions, and through a synthesis of GenAI, MAS and self-organisation, to explore and evaluate the impact of new SAs on their empowerment. While each individual module executes a relatively simple process (and will be enhanced in further work), the collective assembly of them forms an end-to-end functional system which displays a rich variety of behaviours and demonstrates the proof of concept.

The primary contributions are threefold. Firstly, it implements an innovative system for the self-organisation of social arrangements and the visualisation of community empowerment. Secondly, it offers a demonstration of *power-sensitive design* (Mertzani and Pitt, 2024), an instance of value-sensitive design in which the qualitative human values being targeted are power and empowerment. Thirdly, it provides a demonstration of institutional interoceptive awareness as a community detects and responds to a situation that is affecting their empowerment.

The significance of this work is that it proposes a multi-component system which combines human social knowledge and expertise, with the creative capacity of GenAI, stemming from the unexpected linkage of diverse knowledge, and the capability to observe emergence of agent-based simulations in MAS. The system assists humans to think out-of-the-box and

envision future trajectories of alternative solutions, which enables them to effectively self-determine their social arrangements and maintain an equitable distribution of power. This ‘tool’ could support deliberative assemblies, such as humans sharing an office, or stakeholders of a co-housing project, to shape their social arrangements such that they improve their lives.

REFERENCES

- Adams, R. (2008). *Empowerment, participation and social work*. New York, NY: Palgrave Macmillan.
- Bookchin, M. (2004). *Post-Scarcity Anarchism*. AK Press.
- Bouschery, S. G., Blazevec, V., and Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, 40(2):139–153.
- Graeber, D. and Wengrow, D. (2021). *The Dawn of Everything: A New History of Humanity*. London, UK: Allen Lane.
- Jones, A. and Sergot, M. (1996). A formal characterisation of institutionalised power. *Journal of the IGPL*, 4(3):427–443.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
- Kierner, S., Kucharski, J., and Kierner, Z. (2023). Taxonomy of hybrid architectures involving rule-based reasoning and machine learning in clinical decision systems: A scoping review. *J Biomed Inform.*, 144:104428.
- Lewis, P. R., Marsh, S., and Pitt, J. (2021). AI vs ‘AI’: Synthetic minds or speech acts. *IEEE Tech. Soc. Mag.*, 40(2):6–13.
- Manville, B. and Ober, J. (2019). In search of Democracy 4.0: Is democracy as we know it destined to die? *IEEE Tech. Soc. Mag.*, 38(1):32–42.
- Memmert, L. and Tavanapour, N. (2023). Towards human-ai-collaboration in brainstorming: Empirical insights into the perception of working with a generative AI. In Aanestad, M., Klein, S., Tarafdar, M., Han, S., Laumer, S., and Ramos, I., editors, *31st European Conference on Information Systems (ECIS)*.
- Mertzani, A. and Pitt, J. (2024). Power-sensitive design using higher-order cybernetics patterns. In *Artificial Life Conference Proceedings 36*, volume 2024, page 126. Cambridge, MA: MIT Press.
- 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.
- Metz, C. (2016). In two moves, AlphaGo and Lee Sedol redefined the future. In *Wired* (2016), available at <https://tinyurl.com/twomoves>.
- 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*. Springer.
- OpenAI (2024). ChatGPT API. <https://chat.openai.com>.
- Pitt, J., Busquets, D., and Riveret, R. (2013). Procedural justice and ‘fitness for purpose’ of self-organising electronic institutions. In *Proc. PRIMA*, volume 8291 of *LNCS*, pages 260–275. Cham, CH: Springer.
- Pitt, J. and Nowak, A. (2014). Collective awareness and the new institution science. In Pitt, J., editor, *The Computer After Me*, chapter 12, pages 207–218. London, UK: ICPress.
- Pitt, S., Lacey, M., Scaife, E., and Pitt, J. (2021). No app is an island: Collective action and sustainable development goal-sensitive design. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6:24–33.
- Prettejohn, B., Berryman, M., and McDonnell, M. (2011). Methods for generating complex networks with selected structural properties for simulations: A review and tutorial for neuroscientists. *Frontiers in computational neuroscience*, 5:11.
- Robbins, J. (2019). If technology is a parasite masquerading as a symbiont, are we the host? *IEEE Tech. Soc. Mag.*, 38(3):24–33.
- Robbins, J. (2022). The intelligence factor: Technology and the missing link. *IEEE Tech. Soc. Mag.*, 41(1):82–93.
- 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.
- Santos, M. S. and Pitt, J. (2014). Emotions and norms in shared spaces. In Balke, T., Dignum, F., van Riemsdijk, M. B., and Chopra, A. K., editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems IX*, pages 157–176. Cham, CH: Springer.
- Sarkadi, S. (2024). Self-governing hybrid societies and deception. *ACM Trans. Auton. Adapt. Syst.*, 19(2):9:1–24.
- Seidman, A. and Seidman, R. (2009). ILTAM: Drafting evidence-based legislation for democratic social change. *BUL Rev.*, 89:435.
- Thomaz, A. and Breazeal, C. (2006). Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *AAAI Conf. on Artificial Intelligence*.
- Wach, F., Karbach, J., Ruffing, S., Brünken, R., and Spinath, F. (2016). University students’ satisfaction with their academic studies: Personality and motivation matter. *Frontiers in Psychology*, 7:1–55.
- Zarkadakis, G. (2020). *Cyber Republic: Reinventing Democracy in the Age of Intelligent Machines*. Cambridge, MA: MIT Press.