Open-Ended Institutional Adaptation through Human-AI Co-Production

Stefan Sarkadi[1], Anuschka Schmitt[2], Alison R. Panisson[3], Asimina Mertzani[4], and Jeremy Pitt [4]

[1]King's College London, UK
[2]London School of Economics and Political Science, UK
[3]Federal University of Santa Catarina, Brazil
[4] Imperial College London, UK
stefan.sarkadi@kcl.ac.uk; a.schmitt2@lse.ac.uk; alison.panisson@ufsc.br; j.pitt@imperial.ac.uk

Abstract

The ubiquity of Artificial Intelligence (AI), and human interaction with AI, has raised legitimate concerns about the need for and the preservation of human agency, collaboration, and social participation. This development stands in contrast with the collective action and decision-making required to tackle grand societal challenges pertaining a multitude of individuals and stakeholder groups. In this paper, we explore the potential of multi-agent systems (MAS) to increase self-organizing capabilities and participatory decision-making processes using the example of social arrangements. To enable these collaborative processes, our paper makes two key arguments: one, issues of scale associated with closed specification spaces currently limit self-improvement needed for collective organizing. Two, explicitly considering how to design human-AI interaction as open-ended offers a path to address these challenges.

Introduction

The importance of and need for coordination and co-creation become evident if we consider grand challenges such as climate change or digital transformation (Ostrom, 2017), but also micro- and meso-level goals such as organizational learning (Boland et al., 1994). Organizational contexts are defined by increasing complexity and interrelatedness, as well as ever-changing problems, environments, and constellation of agents (Bostelmann-Arp et al., 2022; Parkar et al., 2024; Wan, 2023). At the same time, human-machine systems within organizations, and the design of such systems, appear to undermine humans in their agency and participation, ultimately limiting dynamic adaption and reciprocal interaction among all relevant stakeholders (Parker and Grote, 2022).

Social arrangements, including social contracts, offer important means to self-governance by establishing deliberative and participatory processes that foreground value-driven and contextually aware decision-making (Pitt et al., 2013; Graeber and Wengrow, 2021). Yet, the design and realization of collective organizing, decision-making, and collaboration are not as straightforward. Taking closed specification spaces as a point of departure, we identify three key challenges of scale associated with organizations: the number of members, the longevity of the organization, and competing organizations. As closed specification spaces, e.g., with rule-based systems, are predefined, newly emerging (variances in) priorities, values, problems, and competition for control, cannot be considered. As such, closed specification spaces appear limited in scope for self-improvement.

Human-machine social interactions are arguably harmful, or at least not required, to coordination and co-creation for the grand challenges of today's society. In this paper, however, we assume these interactions as omnipresent and inevitable to our society (Pedreschi et al., 2025). This becomes evident when considering the enforced adoption of AI systems, e.g., in organizational contexts (Bannon, 2023). Humans and machines not only co-exist yet offer the potential to adapt, respond, and learn from each other (Zagalsky et al., 2021). In response, this paper explores how we need to *design* human-machine interactions to aid and complement humans in collective action and collaboration. As such, human-AI co-production as a socio-technical system is explored as a promising, but not the only, path to enhance the self-organizing capabilities needed for collective organizing and decision-making.

In this body of work, we propose to move from closed towards open specification spaces in human-AI interaction to address the complexities associated with self-governance in dynamic organizational environments. Using Socially Guided Machine Learning as a concrete instantiation of dynamic innovation in social arrangements, we explore how Large Language Model (LLM)-enabled Multi Agent Systems (MAS) can serve as promising tools for dynamic self-organization.

'Closed' Institutional Adaptation

This section looks at two forms of institutional specification: firstly using the framework of dynamic norm-governed systems (Artikis, 2009; Artikis et al., 2009), and secondly using social contracts as an equivalent and equally effective shortcut for potentially time-consuming rules-based deliberation and decision-making (Scott et al., 2024). However, it concludes that due to various issues of scale, both approaches

are effectively 'closed', in the sense that there is no scope for further institutional self-improvement.

Social Arrangements

One formal model of self-organising electronic institutions, based on Ostrom's theory of institutions for sustainable common-pool resource management (Ostrom, 1990), defined such an institution as a set of rules, where each rule had a number of changeable parameters, each with a domain of possible values. For example, a self-determination rule involving a consultative vote could have one changeable parameter for the voting method (plurality, Borda count, run-off, etc.), and another changeable parameter to determine the quora (the percentage of enfranchised voters to have voted to qualify for a valid ballot: 50%, 60%, etc.).

This could formally be represented as:

$$\begin{aligned}\mathcal{R} \quad = \quad & (V_{1,1} \times V_{1,2} \times \ldots \times V_{1,P_1}) \times \\ & (V_{2,1} \times V_{2,2} \times \ldots \times V_{2,P_2}) \times \ldots \times \\ & (V_{N,1} \times V_{N,2} \times \ldots \times V_{N,P_N})\end{aligned}$$

where N is the number of rules in the set, $V_{i,j}$ is the number of values that the jth parameter of rule i can take, and P_i is the number of parameters of rule i. $\mathcal{R}$ can be thought as defining a number of *degrees of freedom* for the institution.

This formal representation can be visualised as a *specification space*. A particular point in the specification space represents a single complete dynamic specification – a *specification instance* – and can be defined by an n-tuple. Each element of the tuple is a rule, and is itself a P_i-tuple, where now the jth element is the value of the corresponding parameter P_i.

Given the appropriate metrics, a difference or 'distance' between one specification instance and another can be computed. Then the institution can also specify its own meta-rules about "moving" in the specification space to evaluate proposals for adaptation. For example, certain configurations of parameter values may be considered unacceptable (invalid), or there may be constraints on how "far" the specification can be changed, based on a distance metric d. Figure 1 depicts a specification with (an unlikely) three degrees of freedom, the bold circle indicates the current specification instance, filled circles are invalid, allowable changes are within the gray area. The enactment of proposals (for transitions to alternative specification instances) that do not meet the distance criteria are also invalid. However, these meta-rules are also conventional and could be changed to make these instances accessible.

Social Contracts

In principle, social arrangements support self-governance through deliberative processes, whereby those who are affected by the arrangements participate in their selection, modification and enforcement. The process of self-determination is then concerned with the assignment of values to parameters to define specific configurations of the social arrangements to be congruent with, or fit-for-purpose for, prevailing environmental conditions (Pitt et al., 2013; Graeber and Wengrow, 2021). However, in practice, these social arrangements may need to be applied with a frequency, and within implicit cost constraints, such that performance becomes a pressing issue, and especially so in the presence of existential threats. Moreover, the size of the institution increases, in terms of the number of constituent members, agitates a fundamental tension between the search for consensus (ideal, but potentially impractical) and majority decision-making (practical, but risks majoritarian tyranny) (Mertzani et al., 2023).

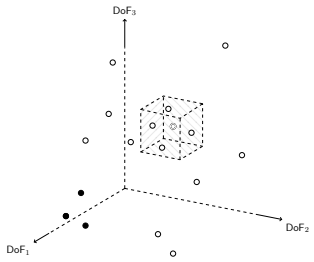

Figure 1: Specification Space with 3 DoF, distance, and invalid instances

A typical way to address this issue is to reduce the number of participants in the deliberative process, either by sortition (which risks exclusion of expertise) or by elective representation (which risks factional control). An alternative to compromise on resource-intensive processes implementing an institutions social arrangements is to replace social deliberation by the use of *social contracts*. This need not be a compromise on democratic processes (cf. (Pitt and Ober, 2018)) any more than or a reduction in participation through sortition or "representative democracy". Instead, the aim of a social contract is to combine an expressive rule representation with efficient and effective rule processing. Essentially, the social contracts converts the specification space of Figure 1 into an 'equivalent' matrix representation that can be efficiently processed (Scott et al., 2022, 2024). Such processing is then effective if outcomes produced by the transformation of social arrangements into social contracts are as "acceptable" or "correct" as the actual products of those deliberative decision-making processes would have been.

Issues of Scale

The use of majority decision-making instead of consensus in search of satisfactory compromises is a reflection of three inevitable problems of scale. The first problem is that as the number of members of the institution increases, so might the variance in priorities, preferences, and values increase. However, within the confines of a predefined specification

space, once the factions of an institution have optimized and counter-optimized against each other, the potential for adaptation stagnates. In other words, the institutional factions have effectively "played all the games" against each other within the 'closed' specification space, and eventually there are no new "games" to discover and play. Having eventually tried every option, the institution has no scope for self-improvement.

The second problem of scale is that the institution itself might encounter other institutions, that are competing for control over the same space, or for the same resources. This additional complexity was studied extensively by Ostrom (Ostrom, 1990), based on which the the requirement for *polycentric* governance was developed. Polycentricity explicitly takes into account the existence of multiple stakeholders, and so multiple autonomous decision-making authorities. These authorities then coordinate their actions with respect to each other. This demands the development of a *polity* (van den Hoogen, 2024; Boomgaarden et al., 2011), i.e., a prescription of 'foreign policy' for dealing with external actors, and distinguishes between politics (small 'p') and statecraft (i.e., intra- and inter-institutional self-governance).

The third and crucial problem is that as the longevity of the institution increases, so might the variance in and the environment in which it is embedded increases, creating problems which have hitherto not been encountered. As before, within the confines of a predefined specification space, once the institution has optimized and counter-optimized against the environment, the potential for adaptation stagnates. In other words (and at risk of repetition), the institution has effectively "played all the games" against the environment within the 'closed' specification space, and again there are no new "games" to discover and play. Having exhausted the set of available options, the institution once again has no scope for self-improvement.

'Open' Institutional Adaptation

To overcome these issues of scale, and address further challenges, this section argues that in contrast to 'closed' institutional specification, what is required is an approach to self-governance in human-AI that promotes 'open', or 'open ended' institutional adaptation. It starts first by looking at organisational modelling in multi-agent systems that enables software agents to represent and reason about institutions, and then proposes the Socially-Guided Reasoning and Learning (SGRL) architecture as way to enable humans and AI to represent and reason about institutions and social arrangements such that it levergages the 'best' of both (e.g. human intuition and inspiration, AI ability to link diverse knowledge). The critical component here is the interactive interface, and we conclude with a discussion of interaction and interaction narratives in the pursuit of open-ended institutional adaptation.

Organisational Models

In contrast to classical multi-agent organisation models that rely on fixed or pre-specified norms, frameworks like JaCaMo (Boissier et al., 2013) offers an illustrative point of reference. Within JaCaMo, the MOISE (Hübner et al., 2006) dimension provides an explicit organisational layer that separates structural, functional, and normative concerns. Roles, missions, groups, and deontic rules are represented in a machine-readable form, enabling agents not only to reason about their organisational commitments but also to coordinate under enforced institutional constraints. This design provides an intersection between explicit institutional specification and practical agent execution, situating it as a frameworks that assume a well-defined specification space.

From the perspective of our argument, this organisational exemplifies both the promise and the limitation of current institutional models. On one hand, MOISE demonstrates the benefits of embedding institutional abstractions directly into agent reasoning and system execution, providing a mechanism for aligning autonomy with coordination. On the other hand, it relies on explicit specifications, highlighting the closure of the institutional design space: norms, roles, and missions must be defined in advance, and adaptation takes place within those predefined boundaries. In contrast, the concept of open-ended institutional adaptation that we introduce in this Section requires that not only the values of parameters or the application of rules can evolve, but also that the institutional representation itself can be re-imagined, re-negotiated, and co-produced through interaction between humans and artificial agents.

This approach provides a precise locus for institutional representation and enforcement that is essential for any adaptive process, yet it remains bounded by a closed specification space. Extending such frameworks towards open-ended adaptation would demand mechanisms for runtime meta-level operations, where new roles, missions, and even forms of normative expression can emerge dynamically rather than being fixed at design time. JaCaMo offers a concrete foundation for institutional modeling, but our proposal pushes beyond it, seeking to integrate continual co-production and expansion of institutional forms as a core capability for human–machine ecosystems.

Innovation of Social Arrangements

Offering this core capability with respect to the identified problems of scale, the effectively closed system of a specification space for social deliberation and its alternative matrix representation for social contracts is insufficient. Instead, what is required is a process that enables dynamic *innovation* of social arrangements.

One approach to this problem it to extend and enhance the Socially-Guided Machine Learning (SGML) methodology (Thomaz, 2006) towards SGRL (Socially-Guided Reasoning and Learning) (Mertzani and Pitt, 2024). SGRL has

a specific application of trying to understand the impact of social arrangements on community empowerment. In particular, though, SGRL tries to leverage the abilities of, on the one hand, human inspiration and imagination, and, on the other hand, Generative AI's capability to link diverse knowledge, to produce *innovative* social arrangements which expand a specification space, or alternatively enhance social contracts with new terms and conditions.

In a variation of the model-view-controller pattern, SGRL uses a multi-agent simulation (MAS) model, a visualization of the simulation (specifically the impact of social arrangements on empowerment), and *two* controllers: the human user and GenAI, as illustrated in Figure 2. Here, the system iterates through a first phase in which the multi-agent simulation visualizes 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. Either the human user or the GenAI can recommend alternative social arrangements: the effect of these new SAs is simulated in the MAS, and the impact on community empowerment is visualised for human 'consumption'. A typical dialogue between human and GenAI in a system for innovating social arrangements that enhance empowerment and community health is shown in Figure 3.

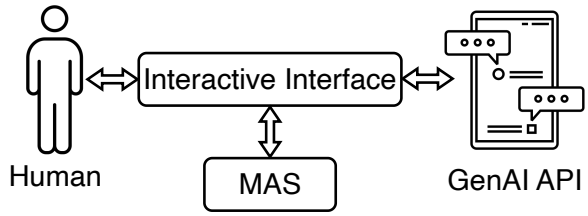

Figure 2: Socially-Guided Reasoning and Learning (SGRL)

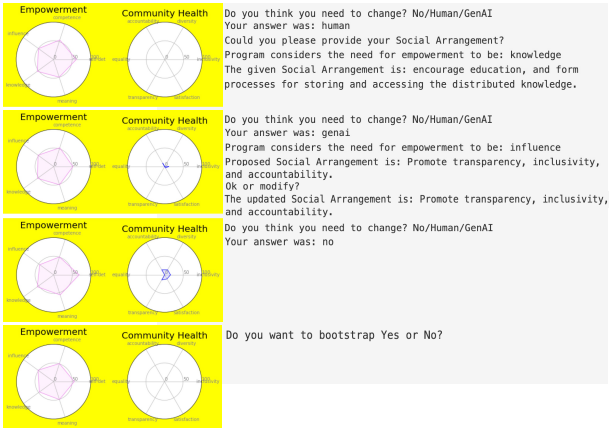

Figure 3: Dialogue between human and GenAI about innovating social arrangements to improve empowerment and community health

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. Moreover, resetting (or bootstrapping) the system allows exploration of multiple different iterations of proposed social arrangements, 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 innovative social arrangements. This integration of human and computational intelligences aims to confine the weaknesses of GenAI (e.g. bias, hallucination), while benefiting from its strengths to support human creativity in avoiding the inevitable entropic effects of 'closed' institutional adaptation.

Rethinking Interaction and Computational Output

At the core of SGRL stands the interactive interface that enables humans to interact in a conversational dialogue (see Figure 2). Next to defining *what* output (i.e., final state) is generated by the system, it is crucial to consider *how* this output is framed (Zagalsky et al., 2021). If we care about human agency and an equitable distribution of power, it becomes questionable if prevalent designs of computational output enable, or rather prevent, such distribution.

Default designs of computational output oftentimes generate solution-focused, unequivocal recommendations. Such designs allow human users to forgo their agency by limiting their efforts to simply confirming or rejecting the output (Miller, 2023). Returning to earlier mentioned challenges of bias and hallucination in output, the design of solution-focused output can become particularly harmful in the context of GenAI-based systems. More particularly, the design of GenAI-generated output defines, and thereby constrains, which solution or option(s) are presented, or made prominent, to the user. This is done by including selected, but excluding other, information and content. Similar to a closed specification space, or algorithmic governance, the design of output thereby constrains, and fixates, what content and information the human user even considers (Pitt et al., 2025). An extensive body of literature on human-computer interaction illustrates how such default, prevalent designs of computational output lead to increasingly converging harmful and unconsidered interaction patterns, including over-reliance, fixation on computational output, and unquestioning delegation (Vaccaro et al., 2024).

Thus, next to envisioning the system architecture and implementation of the interplay among the MAS and the two controllers, one can think more explicitly about the design of the interaction between the two controllers, i.e., human and GenAI. Congruent with this work's main line of argument, i.e., the opening of specification spaces in human-AI interaction, we can also 'open up' the output of a system that is provided to the human user.

In the SGRL context, specifically, computational output serves two key purposes. One, to communicate a final state (of a social arrangement) to the user and, two, to prompt the user to evaluate the generated state and to suggest changes. By 'opening up' the output we mean that, instead of providing a closed, conclusive recommendation, e.g., of a final state, the system first necessitates the input of the human user and rather provides an open-ended, reflection-provoking critique of the human user's input. In addition, computational output can be leveraged to extend the human user's considerations and introduce information previously unconsidered. With both these mechanisms, i.e., challenging and broadening the user's solution space, computational output cannot exist independently but can be made sense of only in tandem with the human user's input and knowledge. In the context of SGRL, specific examples of such output design could entail open-ended questions that might spark the human user to think of new final states or unconsidered aspects key to the social arrangement at hand, or arguments that provide feedback on a user's previous input.[1] Computational output thereby enables inspiration and reflection, or triggers new ideas or actions, yet it is the human user who is embedded in the wider (social) context and who has to recognize the output's meaning for a social arrangement.

Relatedly, the concept of *interactive narratives*, where a user intervenes in or manipulates a (fictional) state, enables and requires the human user to engage in deliberate actions that have fundamental and meaningful implications for the (further) development of a social arrangement (Riedl and Bulitko, 2013). Enabled by the underlying MAS and GenAI-supported interaction, the SGRL system can suggest an updated (or multiple, pluralistic alternatives for a) constellation of a social arrangement based on the user's input or action.

This type of open-ended, reciprocal design of computational output in human-AI interaction serves two key purposes central to the overall aim of open-ended institutional adaptation: first, human power and agency in the participation of a social arrangement is ensured. By design, human users cannot forgo their agency as the computational output cannot be made sense of on its own. In fact, the usefulness of the computational output depends on the quality and extent of the human user's input. This also has the implication of reducing the dimensions of hallucinated output by the GenAI. Second, this open-ended design embodies the essence of sound coordination and co-creation for social arrangements. Rather than seeing human-AI interaction as an exogenously defined, one-time exchange of a computational output and a human user's input, the GenAI-enabled, dialogue-based interaction enables a reciprocal interaction for continuous discussion and deliberation where different inputs feed off each other and can be made sense of only in sequential interaction with each other. Without significant human action or intervention (i.e., beyond accepting or rejecting a computational output), a current state or interaction outcome cannot be altered.

[1] For details and examples of this design, see (Schmitt, 2024).

Challenges Ahead

The difference between Open and Closed MAS does not just rely on issues of scale and the ability of agents to leave or join networks of other agents. More importantly, it relies on the property of interoperability at multiple levels of abstraction. Open MAS is about either one or all of the following properties: agents should be allowed to 1) revise, change, adapt their own internal beliefs and cognitive processes; 2) change the rules of interaction between and withing groups; 3) change/adapt/evolve the very mechanisms responsible for 1) and 2). According to (Sarkadi et al., 2022; Sarkadi and Gandon, 2023), in order to leverage sustainable interoperation in Open MAS, AI designer, engineers, as well as stakeholders need to invest and incentivize translation-based communication between Semantic Web Ontologies, i.e., hypermedia communication, rather than focus on the myth of a 'Universal Ontology'.

From this perspective, there are a variety of challenges for the transition from 'closed' institutional specification to open-ended institutional adaptation, including those examined in this section: autoformalisation and auto*in*formalisation, power dynamics and human-machine semiotics, and the *social* economy of human-AI ecosystems.

Autoformalisation and Autoinformalisation

A foundational aspect of enabling open-ended interactions, such as deliberative dialogues between MAS agents (Parsons et al., 2007; McBurney and Parsons, 2002) is giving machines the capability to translate natural language into formal/computable representations of utterances, i.e., of speech acts (Smith, 2003). Recent work has focused on automatically extracting arguments from natural language text and translating these into computable representations that correspond to Agent Oriented Programming Languages (Trajano et al., 2024). Being able to do this allows AI agents to first execute inferences, check and update their knowledge bases, and subsequently perform actions based on the operational semantics derived form the knowledge base.

Even if current technologies such as LLMs aren't look-up tables such as Searle's description of the Chinese Room (Searle, 1982), they are still reactive agents in a sense. The reaction function operates now on an absolutely gargantuan domain space (like a markov blanket of text embeddings) rather than `string input` $\rightarrow$ `string output` function based on predefined mappings. There are no such things as signs and signifiers in the internal language pattern generation of the LLM.

The open challenge that still remains here is that of Autoinformalisation, that is the translation from for-

mal/computable representations of speech acts into natural language. One can say that by having just a one-way channel human to machine, but not vice versa, means that we cannot consider a machine to actually have the ability to perform a speech act when interacting with a human.

Solving this second challenge will address this interactional imbalance between humans and machines. It will also completely changes the way in which MAS engineering research could being done—with a prominent focus on the human, through enhanced interaction, interpretability and explainability. The key advancement in AI here would be addressing conceptual semantics (Jackendoff, 2006), rather than linguistic semantics.

Power Dynamics and Human-Machine Semiotics

We cannot achieve sustainable Open MAS without considering aspects of resilience. One such important aspect relates to efficient communication in MAS, namely that of self-regularisation mechanisms, which should ideally be implemented to avoid interoperability breakdown. This can be achieved by controlling the influence some agents can exert over others due to power imbalance (Piazza et al., 2025). On the same note, regulation is also crucial for managing deception in Open MAS, ideally through self-governance rather than top-down or pre-defined rules (Sarkadi, 2024).

Regulating the power dynamics of communication and interaction is not enough though. We need to look further into balancing human-machine co-existence by considering the revision of the cognitive (of the individual agents), ontological, and normative realms. Open MAS should allow for the jailbreaking of Human-in-the-Loop systems when such systems do not serve a meaningful purpose anymore. One way to look at human-machine interaction in complex systems is as a cycle and stages of ritual: from birth or emergence of a process of communication that establishes shared knowledge between human and machine (usually due to some fitness property achieved through evolutionary selection), to the enactment of the process by other agents, which becomes a ritual adopted by other agents to communicate meaning through signs and signifiers (consolidation of semiotics through social learning) (Eco, 1979), and, finally, the ossification of the process, which is a ritual that has lost any meaning that is repeated by other agents over generations. In this final stage of ossification agents engage in the interactions without actually knowing why. It is in this final stage where the signals of the semiotic process remain, yet the signifiers degrade. Agents do not have access to the truth-value interpretation model, i.e., the conceptual semantics (Jackendoff, 2006) behind the syntactic signals, nor do they understand the purpose of their interactions.

Social Economy of Human-AI Ecosystems

How can socio-economics, as a paradigm, be applied to the study these cybernetic systems? The main issue, at least from a methodological perspective is that heterogeneous MAS interactions are very difficult to study from a macroeconomics perspective. The usual procedures derived from Game Theory and Mechanism Design are used to 'solve' interactions by applying the *concept solutions*. Yet, these concept solutions fail to actually represent the problem at hand, never mind enabling us to study the ever evolving dynamics between processes.

Indeed, Evolutionary Game Theory does give us some insights with regard to closed dynamics. However, to get any of the EGT insights, the underlying parameters must be predefined, as well as the MAS interaction rules. How do we address these theoretical, and methodological gaps?

From the theoretical point of view, an alternative to Game Theory and Mechanism Design would be Drama Theory (Howard et al., 1993; Howard, 1994).[2].

An exemplification of drama in this sense is the following. Two agents are playing chess, and at some point throughout the game one of the players flips the table—who wins the game? Is there still a game to be won? What happens in such dynamics, is that the agents co-produce new forms and rules of interaction as well as new modes of representing the self, the other, and interdependency relationships.

By extension, this implies that the nature of the reward and payoff mechanisms derived from interactions also change. Imminently, states of equilibria, evolutionary stability and resilience are not anymore guaranteed. Opening MAS systems in this way means opening them up to systemic tragedies —i.e., inescapable events that can be identified through forecasting, but inescapable through backcasting from given states of the system.

To start modelling such interactions, it is also desirable to include the ability of agents (humans or machines) to re-represent themselves as well as others (Lewis and Sarkadi, 2024).

A Debatable Conclusion

To conclude, we summarise this paper using the following dialectic: economists and computer scientist operating under the current MAS paradigm might ask "Why open MAS models in the first place?" and "Aren't our existing concept solutions or dynamical models good enough?". To such questions we give the following reply: *if MAS models are supposed to represent real-world human-machine interactions, then these were open in the first place.*

Acknowledgements

The authors are grateful for the helpful and encouraging comments of the two anonymous reviewers, which have greatly assisted in revising and extending this paper.

[2]We are referring here to Nigel Howard's Drama Theory, not the theories about dramatic arts such as theatre or cinema.

References

Artikis, A. (2009). Dynamic protocols for open agent systems. In Sierra, C., Castelfranchi, C., Decker, K. S., and Sichman, J. S., editors, *8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 1*, pages 97–104. IFAAMAS.

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

Bannon, L. (2023). When AI overrules the nurses caring for you. *The Wall Street Journal*, 15.

Boissier, O., Bordini, R. H., Hübner, J. F., Ricci, A., and Santi, A. (2013). Multi-agent oriented programming with jacamo. *Science of Computer Programming*, 78(6):747–761.

Boland, R. J., Tenkasi, R. V., and Te'eni, D. (1994). Designing information technology to support distributed cognition. *Organization Science*, 5(3):456–475.

Boomgaarden, H. G., Schuck, A. R., Elenbaas, M., and De Vreese, C. H. (2011). Mapping EU attitudes: Conceptual and empirical dimensions of euroscepticism and EU support. *European Union Politics*, 12(2):241–266.

Bostelmann-Arp, L., Mostaghim, S., Braun, A., and Tüting, T. (2022). Multi-objective evolutionary game theory: A case study in cancer therapy. In *Artificial Life Conference Proceedings 34*, volume 2022, page 20.

Eco, U. (1979). *A theory of semiotics*, volume 217. Indiana University Press.

Graeber, D. and Wengrow, D. (2021). *The dawn of everything: A new history of humanity*. Penguin UK.

Howard, N. (1994). Drama theory and its relation to game theory. part 1: dramatic resolution vs. rational solution. *Group Decision and Negotiation*, 3:187–206.

Howard, N., Bennet, P., Bryant, J., and Bradley, M. (1993). Manifesto for a theory of drama and irrational choice. *Journal of the Operational Research Society*, 44(1):99–103.

Hübner, J. F., Boissier, O., and Sichman, J. S. (2006). Programming mas reorganisation with moise+. In *Dagstuhl Seminar on Foundations and Practice of Programming Multi-Agent Systems*, volume 6261.

Jackendoff, R. (2006). *On conceptual semantics*. De Gruyter Mouton.

Lewis, P. R. and Sarkadi, Ş. (2024). Reflective artificial intelligence. *Minds and Machines*, 34(2):14.

McBurney, P. and Parsons, S. (2002). Dialogue games in multi-agent systems. *Informal Logic*, 22(3).

Mertzani, A., Ober, J., and Pitt, J. (2023). θ-learning: An algorithm for the self-organisation of collective self-governance. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, pages 97–106. IEEE.

Mertzani, A. and Pitt, J. (2024). Social implications of socially-guided machine learning for innovation support. In *2024 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–8. IEEE.

Miller, T. (2023). Explainable AI is dead, long live explainable AI! hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 333–342.

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

Ostrom, E. (2017). Polycentric systems for coping with collective action and global environmental change. In *Global justice*, pages 423–430. Routledge.

Parkar, D., Leyba, K. G., Faerber, R. A., and Daymude, J. J. (2024). Evolving collective behavior in self-organizing particle systems. In *Artificial Life Conference Proceedings 36*, volume 2024, page 36.

Parker, S. K. and Grote, G. (2022). Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. *Applied psychology*, 71(4):1171–1204.

Parsons, S., McBurney, P., Sklar, E., and Wooldridge, M. (2007). On the relevance of utterances in formal inter-agent dialogues. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pages 1–8.

Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A.-L., Dignum, F., Dignum, V., Eliassi-Rad, T., Giannotti, F., Kertész, J., et al. (2025). Human-AI coevolution. *Artificial Intelligence*, 339:104244.

Piazza, N., Karimia, A., Soleymanib, B., Behzadan, V., and Sarkadi, S. (2025). Robust coordination under misaligned communication via power regularization. In *Proceedings of ECAI 2025*.

Pitt, J., Busquets, D., and Riveret, R. (2013). Procedural justice and 'fitness for purpose'of self-organising electronic institutions. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 260–275. Springer.

Pitt, J., Mertzani, A., and Ober, J. (2025). Self-governing systems. *Frontiers in Political Science*, 7:1646734.

Pitt, J. and Ober, J. (2018). Democracy by design: Basic democracy and the self-organisation of collective governance. In *2018 IEEE 12th International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*, pages 20–29. IEEE.

Riedl, M. O. and Bulitko, V. (2013). Interactive narrative: An intelligent systems approach. *Ai Magazine*, 34(1):67–67.

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

Sarkadi, Ş. and Gandon, F. (2023). Interoperable AI for self-organisation. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, pages 86–87. IEEE.

Sarkadi, S., Tettamanzi, A. G., and Gandon, F. (2022). Interoperable AI: Evolutionary race toward sustainable knowledge sharing. *IEEE Internet Computing*, 26(6):25–32.

Schmitt, A. (2024). Ensuring human agency: A design pathway to human-AI interaction. In *ICIS 2024 Proceedings. 6.*

Scott, M., Dubied, M., and Pitt, J. (2022). Social motives and social contracts in cooperative survival games. In *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*, pages 148–166. Springer.

Scott, M., Mertzani, A., Smit, C., Sarkadi, S., and Pitt, J. (2024). Social deliberation vs. social contracts in self-governing voluntary organisations. In Cranefield, S., Nardin, L. G., and Lloyd, N., editors, *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVII - International Workshop COINE*, volume 15398 of *Lecture Notes in Computer Science*, pages 57–75. Springer.

Searle, J. R. (1982). The chinese room revisited. *Behavioral and brain sciences*, 5(2):345–348.

Smith, B. (2003). John searle: From speech acts to social reality. In Smith, B., editor, *John Searle*, pages 1–33. Cambridge University Press.

Thomaz, A. L. (2006). *Socially guided machine learning*. PhD thesis, Massachusetts Institute of Technology.

Trajano, G., Engelmann, D. C., Bordini, R. H., Sarkadi, S., Mumford, J., and Panisson, A. R. (2024). Translating natural language arguments to computational arguments using LLMs. In *Computational Models of Argument*, pages 289–300. IOS Press.

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303.

van den Hoogen, E. (2024). *Popular understandings of the European Union: A meaning-centred mixed-methods study*. PhD thesis, Erasmus University Rotterdam.

Wan, C. (2023). Timescales, levels of organization, and multi-objective agents. In *Artificial Life Conference Proceedings 35*, volume 2023, page 137.

Zagalsky, A., Te'eni, D., Yahav, I., Schwartz, D. G., Silverman, G., Cohen, D., Mann, Y., and Lewinsky, D. (2021). The design of reciprocal learning between human and artificial intelligence. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–36.