frontiers | Frontiers in Political Science

TYPE Original Research
PUBLISHED 15 August 2025
DOI 10.3389/fpos.2025.1646734

OPEN ACCESS

EDITED BY
Ana Campina,
Fernando Pessoa University, Portugal

REVIEWED BY
Ivana Stepanovic,
University of Belgrade, Serbia
Marco Antônio Sousa Alves,
Universidade Federal de Minas Gerais, Brazil

*CORRESPONDENCE
Jeremy Pitt
✉ j.pitt@imperial.ac.uk

RECEIVED 13 June 2025
ACCEPTED 22 July 2025
PUBLISHED 15 August 2025

CITATION
Pitt J, Mertzani A and Ober J (2025)
Self-governing systems.
Front. Polit. Sci. 7:1646734.
doi: 10.3389/fpos.2025.1646734

Self-governing systems

Jeremy Pitt[1]*, Asimina Mertzani[1] and Josiah Ober[2]

[1]Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom, [2]Department of Political Science, Stanford University, Stanford, CA, United States

A key driver in the digital transformation of commercial, educational, organizational and social systems is the increasing footprint of Artificial Intelligence (AI). This is producing a different kind of hybrid socio-technical system, which consists of interacting human and artificial "components." However, beyond the realization of basic *Agentic AI*, AI components are likely to be taking over more advisory, supervisory and administrative roles, especially with respect to human components, and potentially without oversight from some external authority. This is a fundamentally different kind of self-governance—i.e., both operational and constitutional decisions concerning the selection, modification, application and enforcement of *social arrangements*—as a co-production of meaningful interaction between human and artificial intelligences. Using examples, this paper scopes out the identifying features of such self-governing systems, which raise several critical political questions about the kind of human rights that could reasonably be expected in such systems. This includes agency, voluntary association, empowerment, innovation and metrication, as they relate to this profound shift in our understanding of "human-computer interaction" and "human-machine teamwork." Finally, given that self-governing social systems don't tend to persist if they can't adapt to a changing environment or resist entropic decay, we consider the idea of continuous self-improvement as a right in itself, from the perspectives of human factors and user experience, and what this implies for human flourishing and the right to human rights in this new techno-political ecology.

KEYWORDS

self-governance, artificial intelligence, human rights, socio-technical systems, self-improvement

1 Introduction

A key driver in the digital transformation of commercial, educational, organizational and social systems is the increasing footprint of Artificial Intelligence (AI). Broadly, we identify an *AI system* as a software component that makes autonomous choices among available options, using some algorithm based on logical reasoning, reinforcement learning, active inference, statistical optimisation, etc. An AI System will then seek to produce a most-preferred outcome according to some criteria, e.g., a rational self-interested AI will seek to maximize individual utility, a pareto-optimizing AI will seek to maximize collective utility, or will act in the present so as to "maximize future freedom of action" (Wissner-Gross and Freer, 2013). For sure, it would have to be classified as an AI System under the definition of the EU AI Act.

Recently, though, the concept of *Agentic AI* (Hosseini and Seilani, 2025) has attracted increased attention and traction. These are autonomous systems of multiple components that can make decisions and perform tasks without human intervention. The potential benefits of extending such cyber-physical control and decision-support systems to AI-enhanced and enabled socio-technical systems have been widely recognized, with applications in identifying complex behavioral patterns (Andrienko et al., 2024), deliberative assemblies

(Zarkadakis, 2020), citizen science (Roszczyńska-Kurasińska et al., 2023), educational gamification for community energy systems (Bourazeri and Pitt, 2018), and pervasive systems (Zambonelli et al., 2022).

The drive toward Agentic AI is producing an advanced form of socio-technical system, which consists of both human and artificial "components" working together in a hybrid combination of the human and the artificial (Abbas and Munoz, 2021; Domingos et al., 2021; Terrucha et al., 2024). Therefore, a viable socio-technical imaginary concerns the development of hybrid socio-technical systems in which AI components are likely to be taking on—or more pertinently, perhaps, taking *over*—more and more advisory, supervisory and administrative roles, even with respect to human components.

Such progress beyond Agentic AI presents a substantive risk to human agency, empowerment and rights. As pioneering cyberneticist Norbert Wiener wrote in *The Human Use of Human Beings: Cybernetics and Society* (Wiener, 1954):

> *if the human being is condemned and restricted to perform the same function over and over again, he will not even be a good ant, not to mention a good human being* [*ibid.*, p. 52]

> *that such [governance machines] may be used by a human being or a block of human beings to increase their control over the rest of the human race* [*ibid.*, p. 181]

> *what is used as an element in a machine, is in fact an element in the machine* [*ibid.*, p. 185]

Wiener clearly identified that the same technology can have both the beneficial potential to free human beings to pursue individual and collective interests, and a dark side of potential disempowerment and dehumanization. The essential problem, that we are facing when AI is the technology under scrutiny, is that an AI System *is* understandable as an agent but *not* as a person. Therefore AI does *not* have "will" or "volition" and is *not* endowed with rights. This could change if AI achieves "personhood," but it has not yet and may never, and the legal definition would be protracted (cf. organizational or environmental personhood). Nevertheless, the fact that AI may be directly in control over humans, may be perceived by humans "as if" (or they are deceived into thinking "as if") it were a person, or may be acting as proxy for some other humans to have control over those humans, has specific higher-order implications for issues of power, empowerment, human rights and human flourishing. To emphasize, our concern is for the impact of AI on the human condition, particularly human flourishing, and specifically not artificial personhood, robot rights or machine flourishing.

Therefore, the hybridization of socio-technical systems combining human (natural) intelligence and artificial (computational) intelligence, and their extended reach into all aspects of human activity and society, presents a difference of *kind* rather than one of degree. In particular, in the absence of any external authority exercising even limited oversight, this raises fundamental questions of, and for, *self-governance*. Self-governance entails both operational and constitutional decisions concerning the selection, modification, application and enforcement of *social arrangements* (Ostrom, 1990; Ober, 2017; Graeber and Wengrow, 2021). These social arrangements are defined as the set of institutional rules, roles, structures, procedures, norms, etc. that are mutually agreed between otherwise autonomous components, with which they voluntarily comply in order to regulate their own, and each other's, behavior, to their collective prosocial benefit. Being mutually agreed, these social arrangements are socially constructed (Berger and Luckmann, 1966), and so are the co-production of "proper" interaction between human intelligence and artificial intelligence—proper in the sense that human consent to comply with the social arrangements is legitimate, informed and revocable (Pitt, 2022).

Accordingly, after grounding the background to hybrid socio-technical systems with examples in Section 2, this paper scopes out the identifying features of such self-governing systems in Section 3. It then addresses, in Section 4, critical political questions of agency, voluntary association, empowerment, innovation and metrication as they relate to this profound shift in our understanding of "human-computer interaction" or "human-machine teamwork." Finally, given that self-governing social systems don't tend to persist if they can't adapt to a changing environment or resist entropic decay, Section 5 considers the issue of continuous self-improvement and what this implies for human flourishing and human rights in this new techno-political ecology. We summarize and conclude in Section 6, arguing that we need to revisit the intersection of cybernetics, self-governance and political theorizing in order to design and operationalise safely the next generation of hybrid "agentic" socio-technical systems that empower people and communities.

2 Hybrid socio-technical systems

In the early days of human development, social systems were composed entirely of human interactions, relationships, and networks. While some social systems theorists (e.g., Parsons, 1972; Luhmann, 1995; Forrester, 1971) have investigated social systems from different perspectives and have used different methodologies, the common assumption is that social systems comprised networks of humans interacting and collaborating. Moreover, they also considered values, collective behaviors, organizational structures, cultural dimensions, and mechanisms for resilience and recovery from crises from a humanistic perspective.

Technology, as a transformative force, can reshape the ways people interact, inter-relate, work and network together. Therefore, social systems can no longer be investigated alone, without considering the impact of technology, nor can technology be perceived as just a tool that serves society. Instead, society co-exists with technology and, therefore, researchers and practitioners have focused on designing technology that accounts for human needs, behaviors, and values. This has produced the concept of socio-technical systems (Baxter and Sommerville, 2011), which are characterized by the dynamic interaction between humans, organizations, and technology. These systems are

not merely technological artifacts but encompass the social structures, behaviors, and processes that shape how technology is designed, implemented, and used within societal contexts, and how technology shapes society (McCluhan's "the medium is the message").

The development of artificial intelligence (AI) technology, in particular, has revolutionized the socio-technical landscape. As identified in Section 1, the term "AI" encompasses multiple computational paradigms, from expert systems for machine reasoning, through multi-agent systems for distributed machine reasoning, and onto statistical optimisation for machine learning (Dhar, 2024). All these paradigms are are essentially identified and labeled by "AI" in the same way, even though they are distinct in their theoretical foundation, algorithmic specification, information processing, practical application, and potential social impact (Scott and Orlikowski, 2024). Nevertheless, all these different paradigms have contributed to enabling new levels of interaction and collaboration as AI evolves from a tool to an active participant, or *agent*, within a social system, and as much an "agent" as a human. Consequently, when we use the term *agent*, we denote either a human agent with natural intelligence, *or* a software agent, i.e., some form of computational intelligence. Hereafter, when we need to refer specifically to human intelligence, i.e., natural life, we will use the term *NLife*; and to refer specifically to computational intelligence or an AI system, we will also use the term *ALife*, i.e., artificial life (Dorin and Stepney, 2024).

This functional (but not non-normative) indistinction between human agent and software agent, and the integration of natural intelligence with computational intelligence, has led to the emergence of *hybrid socio-technical systems*. These represent a new paradigm where humans and AI-driven technologies coexist and coevolve. Central to this paradigm is the concept of Hybrid Intelligence, which emphasizes the complementary capabilities of humans and AI (Dellermann et al., 2019). Humans contribute creativity, decision-making, and emotional intelligence, while AI offers computational power, efficiency, and scalability. Together, these strengths are combined to augment each other, creating a synergy that ideally enables both humans and AI to achieve outcomes beyond their individual capabilities.

Examples of hybrid socio-technical systems illustrate how human-AI collaboration is already shaping diverse domains. One such example is an intelligent co-working space, where human employees and AI agents interact to manage shared facilities such as lighting and heating. In the early stages, decisions may relate to the initial configuration of workspaces or allocation of shared devices. Over time, AI components begin to learn from the behavior and preferences of human users, proposing adjustments to increase energy efficiency, comfort, or productivity. These suggestions are then evaluated, accepted, or modified by humans through interfaces or social conventions, contributing to the development of shared arrangements (SAs) for (co-)governing the space. In this way, the system evolves through co-produced knowledge, shaped by both human judgment and machine intelligence.

However, open-plan offices as scenes of workplace incivility and conflicts of interest are well-known (Bennett and Robinson, 2003). Equally, in data centers, a task manager agent has a goal to increase throughput, and so wants to turn as many processors as possible on. An energy monitor agent, with the goal to minimize energy consumption, wants to turn as many processors as possible off. Therefore it is not inconceivable that such conflicts and episodes of spiraling incivility should arise in intelligent co-working spaces as well, but between the humans and the AI. For example, an agentic energy monitor might want to turn the lights and heating off to achieve its goals, and the human workers, wanting to be warm and be able to see, might try to turn everything on in order to meet their goals of productivity in comfort.

A similar pattern, on a greater scale, can be observed in eco-villages and circular economy supply chains, where the hybridization of natural and artificial intelligences is central to achieving sustainable outcomes. In eco-villages, residents and AI-enabled infrastructure must make collective decisions regarding the distribution of resources such as energy, water, and space. While humans might define high-level objectives and social norms (e.g., prioritizing renewable energy use or fair resource sharing), AI components support these goals by optimizing resource flows, monitoring environmental indicators, or recommending policy adjustments. In circular economy supply chains, AI agents assist in tracking materials, forecasting demands, and identifying reuse opportunities, while human actors are responsible for evaluating and fine-tuning those recommendations based on ethical and practical dimensions. As in the co-housing example, these hybrid systems require ongoing negotiation between human and nonhuman agents, balancing competing interests and learning from feedback to refine collective rules over time. Across these examples, it becomes evident that hybrid socio-technical systems demand new forms of governance which are rooted in mutual adaptation, distributed authority, and iterative co-design of the social arrangements that sustain them.

Overall, hybrid socio-technical systems underscore the importance of interdisciplinary approaches to system design, human-computer interaction, and operationalisation (Hamann et al., 2016). By leveraging hybrid intelligence, these systems have the potential to not only enhance functionality but also ensure that technological advancements remain aligned with human values and societal needs. However, the need for run-time self-configuration, on-going requirements for negotiation and compromise, the prospect of conflict in the face of competing goals or competition over resources—all in the absence of an over-arching decision-making authority—suggest that this is a new *kind* of socio-technical system: a self-governing system. Moreover, the fundamental question remains, whether or not the AI is working on behalf of the humans within the system, or is a proxy for serving the interests of others outside the system, i.e., a "governance machine" as highlighted by Wiener (1954). Consequently, in the next section, we define the components, critical features and typical political problems that need to be addressed in the management and maintenance of self-governing systems.

3 Self-governing systems

In previous work, e.g., (Pitt, 2021), our focus has been on self-organizing multi-agent systems, whether these were either cyber-physical systems consisting of autonomous software agents, or

socio-technical systems consisting of computer-supported human agents. The two types of system converge in the hybrid socio-technical systems described in the previous section. Our concern in this convergence, though, is more than just the definition of structures, the assignment of roles, and the selection, modification and application of conventional rules. We are now primarily concerned with the acquisition, exercise, transfer and release of power and decision-making authority in systems involving both human and software agents—but still in the absence of any external control. This demands *self-governance*, which in turn implies intentional, value-aligned internal regulation. This positions self-governing systems as addressing a higher-order problem in engineering and operationalising self-organizing socio-technical systems.

A *Self-Governing System* (SGS) consists of a group of otherwise independent and autonomous entities pursuing their collective (public) and personal (private) interests in the context of an over-arching, mutually-agreed, and mutable set of socially-constructed conventional social arrangements, by which is meant:

- social arrangements: any kind of convention, norm, policy, rule, contract or law for prescribing acceptable behavior in the context of the system;
- socially-constructed: a mutually-agreed product of interaction between the entities themselves; and
- conventional: unlike physical laws, the rules are breakable; indeed sometimes they have to be breakable to be mutable (i.e., modifiable) to demonstrate the need for modification or benefits of improvement.

It can be assumed that the entities have established channels for reliable communication (which might be achieved using a common language, or any other mutually understood way of reliable signaling), but not necessarily that they have a fully shared set of congruent values and goals. Moreover, entities cannot be considered identical in competence, and are distributed over a number of different preferences and attributes (e.g. avoidance of tyranny, propensity to cheat, etc.).

Therefore critical features for sustainability of such systems are that the over-arching set of rules should specify and/or identify:

1. Articles of (voluntary) association, for inclusion and exclusion of entities, and expectations of entities in the system (e.g., adherence to the rules, participation in the application of the rules, etc.).
2. Structures and procedures for facilitating the intended functions of the system.
3. Structures and procedures for run-time self-modification of the over-arching set of social arrangements.
4. Processes for knowledge aggregation (deliberation and decision-making) and knowledge alignment (collective action in the public interest).
5. Trustworthy gatekeepers and processes for reliable knowledge codification.
6. A non-repudiable means of monitoring compliance, reporting non-compliance and resolving disputes, with appealable disciplinary procedures.
7. Limits of and rights to self-determination, whereby the selection, modification, application and enforcement of the rules can be conducted within specified guardrails and with recognized rights.
8. The specific form of political organization defining the polity, in particular determining the organizational behavior with respect to external actors and authorities.

The specification, and in particular the operationalisation, of the over-arching set of rules faces many problems encountered in political science and political philosophy:

- sustainability: features of, and design principles for, self-governing institutions to ensure sustainability of self-organization (Ostrom, 1990);
- fairness and justice: canonical principles of distributive justice expressed as legitimate claims in a sector context to ensure fair allocation of resources (Rescher, 1966; Rawls, 1971);
- knowledge management: processes for knowledge aggregation, alignment and codification to ensure "correctness" in deliberation and decision-making, emergence of expertise through social influence (Nowak et al., 2019), and successful collective action (Ober, 2008);
- legitimate consent: the avoidance of tyranny through voluntary association, negotiation and operationalisation of articles of association, which are informed, meaningful and revocable (Ober, 2017);
- legitimate dissent: the tolerance of dissent when outcomes of current practices diverge from a community's shared set of congruent values (Burth Kurka et al., 2019);
- iron law of oligarchy: resisting the tendency of any organization, no matter how democratically founded, to degenerate into control by a few who run it in their own, rather than the common interest (Michels, 1962 [1911]);
- unrestricted self-modification: the observation that unrestricted self-modification of a set of conventional rules tends to paradoxical rules, inconsistency or incompleteness (Suber, 1990);
- voting paradoxes: standard problems in social choice theory about mapping expressed preferences over ranked candidates into a specific choice or choices (Arrow, 1951; Regenwetter and Grofman, 1998);
- citizenship and rights: for example, the human right to a human decision (Tasioulas, 2023), which is diminished by the increasing automation (using "AI") in classification and selection tasks, e.g., for job selection, policing and sentencing;
- dignity: an individual and collective value, which is reinforced if people are able to make meaningful and valued contributions to issues of public interest, and is undermined if people are deceived into making decisions that they would not have made had they been aware of full or accurate knowledge (Pitt et al., 2020).

However, while these are "old" questions appearing in the "new" context of self-governing socio-technical systems, the integration of human and computational intelligence presents a number of "new," and critical, political questions, as discussed in the next section.

4 Critical political questions

The previous section established the foundational features of self-governance for the hybrid socio-technical systems presented in Section 2. It also examined how the digitalisation of social arrangements encountered several problems familiar to political science. However, as the computational components of the socio-technical system demonstrate increasing intelligence, this raises several critical political questions about the kind of human rights that could reasonably be expected in such systems. These rights specifically relate to:

- agency and constitutional choice;
- voluntary association and affinity;
- empowerment;
- innovation; and
- metrication.

This section discusses each of these critical questions in turn. Many of these questions raise the issue of value pluralism, i.e., determining what sort of rights and restrictions a self-governing system could possibly, or even reasonably, put on a human or a computational intelligence's scope for action in a situation in which not all values can be simultaneously accommodated. This implies understanding what it even means for an AI component to represent and reason with "its" values, and indeed, whose values are being represented and reasoned with.

4.1 Agency and constitutional choice

In the early days of text editors and windows interfaces, a frequent user complaint was being "moded in": the computer program controlled the mode of use, and user action was constrained accordingly. This pattern was subsequently observed in the concept of *algorithmic governance*: the computer decides the space of available options and the human is limited to choosing between those options. As artificial agents increase their participatory footprint in human society, including taking administrative and supervisory roles in SGS, there is risk that the corresponding risk of a loss of agency for humans. With limited opportunity to complain, there is a risk that this will prove to be, literally, dehumanizing.

Therefore, critical questions for hybrid self-governing socio-technical systems are the dimensions and parameters of *agency* (Williams et al., 2021). This concerns both the agency of human intelligence (i.e., natural life, or NLife) and computational intelligence (i.e., artificial life, or ALife), and their respective roles in self-governance. This includes mutual participation and co-existence in social and communal processes and structures.

However, it is not just the nature of agency that is of concern but also the *process* of *agentification*. This is a dual process: firstly; as ALife gains agency, as it moves from being a tool that can augment the performance of NLife, toward being imagined or understood as an independent actor, partner or even stakeholder in its own right; but also secondly as NLife *loses agency*. This can occur if the ALife occupies a supervisor or coordinator role with respect to NLife, and if the ALife is perceived by the NLife as being of superior 'intelligence' or status, then the NLife can transition from having had agency to a situation of dehumanization and disempowerment (Robbins, 2019; Wiener, 1954; Milanovic and Pitt, 2021), especially with regards to constitutional choice.

The process of constitutional choice is not about ensuring that specific individuals can get everything they want; it's about agreeing on a shared framework for compromising over the pursuit of the individuals' often conflicting aims (Manville and Ober, 2023; Mertzani et al., 2023). Democratic regimes have proved highly effective in this regard, especially if a *civic bargain* is maintained—through continual deliberation, negotiation, and compromise. Indeed, dissent in deliberation and compromise in negotiation, which could be considered noise, enable majority rule to be an effective approach to decision-making that avoids majoritarian tyranny. This becomes possible when the underlying principle is consensus achieved through democratic deliberation (Canevaro, 2018; Mertzani et al., 2023), or a fair bargain, struck through agreed upon procedures conducted among agents with potentially competing interests (Manville and Ober, 2023).

However, civic bargaining is conceived as a process conducted among persons who see each other as in some meaningful way as equal sharers in a common enterprise. This raises some difficult questions on the tension between value pluralism and taking as a premise basic "no-boss" democracy—as in the demographic sortition that begins the Demopolis thought experiment (Ober, 2017). The point here is that absent that original sortition, there can be people who *want* an ultimate boss, and therefore reject participation by "non-bosses" as inherently illegitimate. Therefore, value pluralism cannot be unbounded: so long as AI systems are not persons, human values take priority. While agency can appear functionally equivalent, personhood is not, and, for example, a rule about automatic termination for non-compliance would not affect an AI as it would a person.

Currently, ALife is some way from being an equal sharer, and it is perhaps unclear if it would, or should, ever have that status: the distinction between person and tool seems to be still the essential one. However, at the point that ALife becomes a person—however we understand that as a description of an entity with moral standing akin to those we now think of unambiguously as persons—and thus potentially a citizen, it has passed out of the domain of tool. If legal personhood can be bestowed on a commercial organization, it is not unreasonable to imagine that such a status could be attributed to a computational process, i.e., ALife, as well, and arbitrarily appointed as a "citizen" of some kind, with concomitant rights in a self-governing system. Lacking a clear distinction between tool and person, it remains unclear what the role of ALife would rightfully be in processes of constitutional choice, civic bargaining, and the determination of citizenship issues, where ALife intervention in some of these processes might be inherently unsafe too.

4.2 Voluntary association and affinity groups

Kropotkin was one of the late 19th century's most prominent advocates of mutual aid through voluntary association (Kropotkin, 1902). Mutual aid was observed to be a way of resisting the

centralizing force of the emerging modern state, with its increased emphasis on individual competition over collection cooperation, the hollowing-out of communal institutions, and the convergence of both sovereign and bureaucratic power. He pointed to numerous examples of mutual aid through voluntary association, including labor unions, charitable trusts, mutual insurance (the forerunner of building societies rather than banks as a place to save money), and even literary salons and scientific societies (e.g., the UK's Royal Society adopted a motto (*nullius in verba*—"take nobody's word for it") that represented its aversion to the domination of arbitrary role-based authority and preference for empirical experiment over "received wisdom").

In the 20th century, Weil drew a contrast between subjective individual rights and objective mutual obligations, and argued that obligations are the more fundamental concept, on the grounds that these obligations stem from satisfying the vital needs of every human being (Weil, 2002). Inverting Maslow's hierarchy of needs, Weil claimed that being rooted (rather than self-actualisation) is the most important but least recognized need of the human "soul." She further asserted that a human being "grows" these roots through voluntary association with, and active, purposeful participation in, the life of a community. Moreover, like a plant's root systems, roots provided both unity, through stable anchorage, and diversity, through multiple associations of location, kinship, education, workplace, and professional activities. Furthermore, these communal associations offer both a common memory of the past and common expectations for the future, creating meaningful mutual bonds between individuals.

The formation of these bonds have a particular impact on what Bookchin called *affinity groups*, whose members are as much concerned with human relationships as they are with the problems facing the group, or the group's role within a social movement (Bookchin, 2024). Such affinity groups are clearly reflected in sporting association: for example, historically, supporting a local community football club was less about whether the team won or lost than about just being at the match and being part of the conversation afterwards (Starkings and Brett, 2021).

These conceptual processes of voluntary association, creation of roots, and establishment of affinity groups which then validate, propel and reinforce subsequent acts and articles of voluntary association, need to underpinned by the notion of *legitimate consent* (Pitt et al., 2025). Legitimate consent demands that processes of voluntary association should be informed, meaningful and revocable: informed, in that an associate should understand what they are committing to, and what to expect of others within the constraints of the association; meaningful, in that the association should demand some for of active participation, and return some kind of reciprocal benefit as a consequence; and revocable, in that the associate should be able to withdraw at will from the association (there being a substantive difference between voluntary association and indentured servitude).

Given the centrality of voluntary association and the social construction of roots within affinity groups, the critical political questions for self-governing socio-technical systems are, firstly, how to prevent these processes from being diminished by increased computer mediation; and secondly, how to replicate these conceptual processes in the direct voluntary association of NLife with ALife.

The former issue is increasingly occurring in those socio-technical systems involving AI-mediated, computer-mediated, human-human communication. Computer-mediated human-human communication (CMC) is a common by-product of advances in ICT which has resulted in e-commerce, e-health, e-learning, and other domains of activity prefixed by "e-." AI-mediated CMC is a product when the human uses an AI, typically an LLM, to produce the intended communication with another human. However, the convenience and availability of LLM has lowered the barrier to communication and so expanded the volume: for example, using an LLM, it has become much easier for students to email professors, or for job-seekers to apply for open employment positions. To cope with the volume, professors and employers in turn are using LLMs to summarize emails or filter applications. The overall result is increased social distance, ineffective roots and diminished affinity.

The latter issue exposes the tension between the pluralistic values likely to be exhibited between NLife, even (as discussed above) amongst themselves, and the ALife. ALife is unlikely to have any "values" themselves, except those implicitly and indirectly encoded by the developers. This is likely to impinge upon the fundamental nature of interaction and the social construction of digital relational commons, and their ability to promote successful collective action in the Digital Society.[1]

4.3 Empowerment

In a self-governing system, the term *empowerment* refers to the awareness and capability of the individuals in a group to exercise choice and control over their social arrangements, i.e., the set of rules, roles, structures, procedures, policies, norms, conventions, contracts or laws, with which they voluntarily agree to comply, in order to hold each other accountable. For both ALife and NLife, this demands the capacity to represent and reason about five cognitive dimensions: self-determination, competence, influence, knowledge and meaning (see Pitt et al., 2025 for details).

Focusing on self-determination through selection of political regime, there is of course a wide range of choice, as evidenced by the variety of words with *-ocracy* or *-archy* suffixes. However, these can be categorized, as per (Ober, 2017), according to the answer to the question *who rules*? In the context of self-governing systems, one answer could start by considering *how many rulers*? and *in whose interest do they rule*? For the question of *how many rulers*?, an answer could broadly be drawn from three options:

- *one*, i.e., an individual (monarchy, autocracy, etc.); or
- *few*, i.e., a small and exclusive coalition selected according to some specific criteria (aristocracy, oligarchy, etc.); or
- *many*, i.e., an extensive and inclusive body of citizens (democracy).

1 Smit, C., Abbas, R., and Pitt, J. (Submitted). Digital polycentricity for sociotechnical design: outcomes from the 2024 workshop on digital polycentricity. Communications of the Association for Information Systems. *Commun. Assoc. Inf. Syst.*

If, alternatively, the answer is "none of the above," i.e., all decision-making is delegated to some *external authority* outside the system, and which may not be affected by the social arrangements at all, then the system is, by definition, not self-governing.

Of the three valid options, this is not necessarily a preference ranking. Classically, Plato (1974) drew a distinction between a "perfect" form of governance by a knowledgeable elite and various "degenerative" forms, while Aristotle (1981) distinguished between political regimes that served the common interest and those that prioritized the ruler's interests. In line with this latter classification, and motivated by our works on Demopolis (Ober, 2017) and SimDemopolis (Pitt and Ober, 2018), we argue that, within SGS, there are substantively better and worse forms of the rule of one, few, and many. Given the difficulty of establishing and sustaining genuinely consensual forms of the rule of one or a few, our focus here is on democracy (rule by many). Two features of democratic empowerment are then, firstly, not getting "stuck" in a form of governance that is no longer fit-for purpose (Pitt et al., 2015), and secondly, resisting a drift to a worse form (Michels, 1962 [1911]; Bermeo, 2016). This could entail being able to switch between one form and another if circumstances demanded (e.g., a situational crisis), so long as enforceable guardrails are established to prevent the accumulation of, abuse of, or unwillingness to relinquish power. It also entails introspective mechanisms that diminish a supposed "law" of social order into a manageable threat, i.e., degeneration into tyranny or oligarchy poses the same kind of threat to stability and prosperity as security (e.g., a threat from hostile external actors) and insecurity (e.g., an inability to provide basic welfare) (Ober, 2017).

However, the difficulty of evaluating fitness for purpose is again the presence of value pluralism. Specific values may be promoted or demoted within each type of regime, or change between regimes; indeed political deliberation can be seen as a compromise on policies relative to different priorities or preferences on values, grounded in mutually-agreed facts or evidence. A SGS will aim at promoting the "health" (or, per below, "flourishing") of the collective, but, in light of value pluralism, there will be points of disagreement and need for compromise on values bearing on collective health, e.g.:

- *safety*: the priority for any system of governance is the safety of its citizens (as per Cicero); welfare and security are two of the three fundamental provisions of Basic Democracy (Ober, 2017), the third being the avoidance of tyranny;
- *cognitive efficiency*: how much of their cognitive resources do citizens have to expend on matters of political discourse as opposed to other socially productive efforts, see for example the role of social influence in distributed information processing (Nowak et al., 2019);
- *inclusivity*: the extent, in terms of opportunity and actuality, that citizens are engaged in selecting, modifying and enforcing their chosen social arrangements, cf. Ostrom's third principle of self-governing institutions (Ostrom, 1990);
- *participation*: the principle that, as per (Ober, 2017), citizens should participate, and be able to participate, equally in matters of political concern;
- *accountability*: to what extent are decision-makers disproportionate beneficiaries of their decisions, to what extent are they rewarded/punished for correct/incorrect decisions, and to what extent does accountability contribute to systemic self-improvement;
- *dignity*: civic dignity is increased when citizens are treated as equal participants in political processes, and diminished when citizens are tricked into making decisions which they would not have made with knowledge of "the facts" (Ober, 2017); however, dignity must remain a threshold condition, determining by "how much" it may have been improved or diminished by political action remains obscure (Hitlin and Andersson, 2023).

4.4 Innovation

As hybrid socio-technical systems continue to emerge, the question of innovation becomes not merely a matter of technical progress or creative output, but a fundamentally political concern. In this context, innovation refers to the collective capacity to reflect upon, generate, and apply novel SAs to sustain and self-improve. In systems comprising both NLife and ALife, where power asymmetries, epistemic pluralism, and value conflicts are inherent, innovation becomes a necessity. This is because it is a mechanism through which communities can reconfigure the foundations of coexistence, challenge the status-quo and envision alternative SAs that can lead to preferable (i.e. more fair, effective, or resilient) societal trajectories. As such, innovation does not simply aim to solve predefined problems; it opens space for posing new questions, constructing new forms of agency, and articulating new principles for shared governance.

Crucially, innovation in hybrid socio-technical self-governing systems is not a one-off event, nor does it occur in isolation. It is a recursive process of reflection and learning which is shaped by feedback loops. To maximize the effectiveness of these processes, NLife and ALife need to work together in such a way that they complement each other. Specifically, effective innovation requires bringing together the capability of ALife to process large-volumes of data and perform evidence-based inference, with the expertise and lived experience of NLife which allows performing value-based assessments. As demonstrated through the Innovation Support System through Deliberation (Mertzani and Pitt, 2024), innovation can be operationalised as an iterative and co-produced process, in which human users and artificial agents collaboratively explore, simulate, and refine alternative SAs. This iterative cycle does not just test the functional adequacy of new SAs; it also engages with their ethical, political, and epistemic dimensions. The innovations that emerge are thus not externally imposed or pre-validated by abstract metrics but are instead the product of situated deliberation, interpreted through local values, contested meanings, and shared aspirations.

From that perspective, innovation functions as a modality of political agency, extending beyond the reactive adjustment to system dynamics into the proactive shaping of institutional and normative frameworks. It is a practice of criticizing current practices, imagining alternatives and asserting the legitimacy to

do so. As such, innovation is interconnected with empowerment: while empowerment enables participation in the formation of SAs, innovation expands the space of what can be considered, imagined, and realized. The reflexivity that underpins innovation is epistemic and political; it requires recognizing that current arrangements are temporary, revealing hidden assumptions, and allowing room for dissent, diversity, and future change. Hence, innovation becomes not merely an instrument of system optimisation but a condition for civic imagination and self-determination within complex hybrid socio-technical ecosystems.

However, capacities for innovation, such as empowerment, are unequally distributed and politically contested. Who gets to innovate, under what conditions, and with which tools, are questions that cannot be ignored. Innovation processes risk reproducing domination or exclusion unless they are designed with procedural sensitivity, with embedded mechanisms for inclusivity, transparency, contestation, and responsiveness. Without such safeguards, deliberative mechanisms can degenerate into technocratic gatekeeping, and innovation itself can become a form of epistemic enclosure. Therefore, the design of innovation-support systems must attend not only to their computational efficiency or output quality, but to their capacity to foster equitable agency, to recognize plural ways of knowing, and to enable sustained civic engagement which recognizes diversity.

Accordingly, innovation must be treated not only as a design problem but as an ongoing political question; a space in which power is shared, negotiated, and redefined. It is essential to foreground this political character of innovation: its potential to either democratize or disempower, to either include or marginalize, depending on the institutional and technical infrastructures in which it is embedded. In hybrid socio-technical systems, where the boundary between governance and computation is increasingly blurred, the capacity to innovate must remain open, participatory, and contestable. This capacity is what guarantees that self-governance is not only sustained over time, but also remains meaningful; grounded in the ongoing process of redefining what it means to live together better.

4.5 Metrication

One way or another, metrics will affect decisions (Hauser and Katz, 1998), as they provide important information and ease understanding of problems and in most cases constitute criteria for decision-making (Patterson and Miller, 2012). Although metrics play a fundamental role in systemic self-improvement, wrong choices of metrics, misinterpretation and misuse of them are some of the common challenges encountered in cyber-physical institutions.

Initially, metrics are hard to define since they should be linked with the system characteristics. Taking a metric from one system is not guaranteed to provide the same information about another system. Also, metrics should be defined so that they really capture the information that they are supposed to describe. Since metrics affect decisions and actions, it is also important to design them in a way that they capture not only a single parameter but also the side information related with that parameter, which can be accomplished by defining auxiliary or multiple metrics. For instance, if something is affected by a, b, c and d, the choice of a metric that captures only a and b might lead to actions that control a and b, but neglect c and d.

Another issue in metric interpretation is the fact that metrics can be self-referential. The definition of a metric might require the knowledge of the outcome of another metric or the combination of the knowledge of some other metrics. Therefore, metrics are interconnected, and consequently, the analysis of their results in order to take actions should take into consideration the values of all relevant metrics.

Moreover, while the first step toward getting some understanding over the system is to identify the appropriate metrics to describe the present, the next step toward achieving organizational goals, such as sustainability or balanced tensions between different incentives, is to add some *meta*-metrics that describe the rate of change of the system and provide visibility over the intertemporal evolution of the corresponding observations. However, defining *meta*-metrics is challenging, while analyzing and understanding them is even trickier, especially for as an internal observer (e.g., an agent). As a result, in many cases systems fail to adapt and maintain sustainability because the metrics that they use reflect only short-term effects.

While metrics are undoubtedly important, organizations many times become victims of those metrics and that is because they end up being obsessed with metrics instead of focusing on identifying the right metrics that provide them with the desired information. As a result, individuals spend too much time and effort in finding ways to measure performance, and end up in a situation in which they disregard matters of substance. Therefore, it is a challenge to identify the minimum required metrics that provide clearly the desired information, and avoid being a victim of over-metrication.

Although defining a metric is one problem, finding an appropriate or meaningful way to employ it presents another kind of problem. The fact that a set of metrics is defined is not enough to guarantee sustainable self-improvement. The agents of the system need to have access to these metrics, the ability to interpret them and the willingness to adapt their behaviors and policies based on the feedback from applying the metrics.

Finally, to mitigate all the possible issues in metric definition, interpretation and application, metrics should be open to change over time, even if the policy of the system is not modified. This is because, first and foremost, even if the wrong metric is chosen originally, a modified metric might produce the intended information. Additionally, in dynamic institutions of dynamic populations, a change of the metrics is required to capture changes in the agents, changes in their knowledge, and any changes in their needs and practices.

5 Continuous self-improvement of SGS

A self-governing social system is unlikely to persist for any significant period of time, if it cannot adapt to a changing environment or create localized order in the midst of entropic decay. For comparison, Ostrom's self-governing institutions persisted over multiple generations, even being maintained by

generations who were not party to the original formulation of the institutional rules (Ostrom, 1990). This is an issue of continuous self-improvement (Bellman et al., 2014): not necessarily self-improvement as an absolute measure and monotonic requirement, but continuous self-improvement relative to changing requirements and operating conditions. This should not affect rights: rights, if they are to mean anything, cannot just be taken away because the environment has somehow shifted. This section examines continuous self-improvement of self-governing systems from two perspectives: firstly from the perspective of human factors and user experience (UX), and secondly from the perspective of human flourishing and human rights in general (i.e., the right to rights).

5.1 Human factors and user experience

Human Factors and User Experience can be defined and distinguished as follows:

- Human factors: the application of psychological and physiological principles to the engineering and design of products, processes, and systems; and
- User experience: understanding how a user interacts with and experiences a product, system or service.

In the case of SGS with respect to the critical political questions in pursuit of continuous self-improvement, we are therefore concerned firstly with the application of psychological principles in the design of political structures and processes, and secondly to the psychological impact on those affected by the delivery of political structures and processes.

5.1.1 Agency

5.1.1.1 Human factors

Value-Sensitive Design (VSD) has been proposed as a design methodology for socio-technical systems, that aims to target specific qualitative human values as higher-level "supra-functional" requirements, beyond the standard functional and non-functional requirements of software systems engineering (Friedman et al., 2008). Extending this, a design framework for self-organizing socio-technical systems has been proposed that complements the VSD methodology with a number of design principles for a core set of critical human values, including: sustainability, sociability, justice, legitimate governance, and prosocial incentives (Pitt et al., 2017). This framework recognizes that one of the key human factors in self-governance is the problem of prosocial incentivisation for equal participation, in the sense of equality of opportunity (Ober, 2017). This is the problem of translating (potential) agency into (kinetic) action, recognizing both the centrality of transactions and reciprocation in the conduct of human affairs, and the importance of non-monetary (qualitative) values attached to those transactions, usually represented in the form of conceptual resources (sometimes known as social capital). Therefore reliable and non-repudiable transactions in different types of non-monetary economy (e.g., reputation, gift, relational, informational, etc.), and the use of these conceptual resources as units of exchange, are essential to increase the social benefits of cooperation and self-governance.

5.1.1.2 User experience

Robbins, quoting Yonck quoting Wissner-Gross, offers a functional definition of intelligence as *Intelligence acts as to maximize future freedom of action* (Robbins, 2022). Under this definition, there is a potential conflict of interest in systems with human intelligence interacting with computational intelligence. Appealing to the second law of thermodynamics, Robbins argues that while intelligence can successfully create goal-directed order out of disorder, which is essentially the outcome of self-governing systems, that order must and will be entropically compensated. He then suggests that what compensates for increased organization, as produced by the co-production of self-governance by human and computational intelligence (that shrinks entropy), is the environment. Robbins insists this is not just the natural environment, but also humans themselves. Not, as Wiener also observes (Wiener, 1954), the outlier human intelligence of a technocratic elite, but "the rest of humanity whose freedom of action …is being increasingly trapped by design" (Robbins, 2022 p. 85). Since it remains possible for even supposedly democratic institutions to be hollowed out (Bermeo, 2016), particular care has to be taken to ensure that increased agency of computational intelligence does not result in diminished agency of human intelligence, especially as a consequence of off-loading or out-sourcing cognition for the sake of convenience.

5.1.2 Voluntary association

5.1.2.1 Human factors

Voluntary association, for the purposes of collective self-governance by citizens, is beset by a *boundary problem*, which is the circularity involved in the definition of citizenship by those who have, at some historical moment, declared themselves to be citizens. Relatedly, it carries the risk of devolving into majoritarian tyranny, marginalizing and harming minority populations and/or those who are stranded outside the body of enfranchised citizens. Even then, as a study of participatory budgeting has shown (Roszczyńska-Kurasińska et al., 2024), decisions and decision-making processes can come to be dominated by a self-empowered minority of privileged members who have the background (education) and the resources (time, energy, social network, etc.) to participate in complex procedures.

5.1.2.2 User experience

As an exemplar of a community event, parkrun is a weekly, volunteer-driven, voluntary-participation physical activity. It is generally less about timing or winning than taking part. One meta-study has shown that social well-being was a primary self-reported factor for regular participation in parkrun, and in particular that future attendance is most strongly correlated with historical attendance (i.e., the more someone participates, the more likely they are to participate) (Peterson et al., 2022). Thus voluntary association with prosocial individual and collective outcomes is a critical factor in achieving *social cohesion*. Social cohesion is one of the most important determinants of successful and sustainable human communities and social systems (Fonseca et al., 2019), yet

coherence appears to be one of the hardest community qualities to define and metricate (Nowak et al., 2019). The mechanics of providing a positive user experience of voluntary association that motivates repeat performance, that is in turn an indicator of social cohesion, needs to be better understood.

5.1.3 Empowerment

5.1.3.1 Human factors

Beyond empowerment as individuals, a community needs to be empowered to reason and function as a *collective* entity. This is especially so when reasoning about, and taking action to preserve or improve, the *health* of the collective. This is akin to an interoceptive sense (like thirst or hunger), so that the human factors involved in empowerment are related to *interoceptive collective awareness* (Pitt and Nowak, 2014). Interoceptive collective awareness is a critical factor in political operationalisation and in deciding whether or not to change social institutions according to prevailing environmental conditions. It may require evaluating the political regime according to specific metrics.

5.1.3.2 User experience

Community complexity (Rychwalska et al., 2021) needs to be matched to task or goal complexity, as the age (indicating skills and experience) and growth (increase or decrease) of a community are both factors that impact a collective's capacity for self-governance. In addition, many local communities are capable of sorting out local problems. However, they would benefit from consultation and access platforms, providing them access to and leverage with external sources of knowledge and funding.

5.1.4 Innovation

5.1.4.1 Human factors

Considering the human factors in the innovation of social arrangements through co-production by interacting NLife and ALife, the crucial issues are imagination and knowledge. On the one hand, it is evident that people do not need to experience directly or have empirical evidence of particular social arrangements or political regimes to believe whether or not these arrangements are preferable to some other alternative arrangement. Imagination, and indeed sometimes intuition, i.e., belief without evidence, can be enough. On the other hand, ALife, especially in the form of Generative AI, can be remarkably effective in linking diverse sources of knowledge in unexpected ways. To leverage these complementary capabilities in the co-production and innovation of social arrangements in the tradition of HABAMABA (humans are best at; machines are best at), while avoiding cognitive biases like automation bias or the human tendency to "dumb down" when confronted by supposedly superior intelligence (Robbins, 2019), is essential for continuous self-improvement of self-governing systems.

5.1.4.2 User experience

There are a number of potentially pernicious interactions between human psychology and the innovation of social arrangements, that can adversely affect the experience and perception of the quality of self-governance. In sociology, for example, *interactional justice* has been defined as the extent to which people affected by the decisions of an institution (or organization, or community) are treated with dignity or respect (Schermerhorn et al., 2011). It has been further refined in organizational theory to include two different forms of interpersonal treatment: one dealing with the extent to which stakeholders in an institution are dealt with by the decision-making executive implementing procedures (called *interpersonal* justice), and the other dealing with the explanations offered to stakeholders about how particular procedures were followed or why certain outcomes were reached (called *informational* justice) (Greenberg and Cropanzano, 1993). A computational treatment of interactional justice used opinion formation over social networks in a multi-agent in order to aggregate a set of subjective, individual opinions into a single collective judgement on the fairness of an institution (Pitt, 2017). In addition, the political philosopher Rawls (1971) tried to evaluate a political regime or institution according to how "fairly" its procedures treat its citizens or members and the development of a "well-ordered society." These two qualifiers of justice together address a fundamental question of dynamical social systems and the innovation of social arrangements: who benefits from constitutional reform?

5.1.5 Metrication

5.1.5.1 Human factors

There are a number of features between human psychology and "measurement" or metrication which can adversely affect the perception of the quality of self-governance. These includes: the tyranny of metrics (Muller, 2018), which is the observation that a fixation on metrics in order to evaluate performance can distort and diminish performance, as people change behavior in order to "max out" the metric rather than advance or achieve organizational goals; Goodhart's Law (Goodhart, 1975), which formalizes the experience that when a measure is used as a target it ceases to carry any meaning; and the quasi-quantum effect on human behavior, where awareness of being measured affects task performance, as does consultation; vanity metrics (Ries, 2011), whereby a metric that appears to be impressive but is relevant only to those whose are impressed by the metric, and is not indicative of true performance, for example *h*-index; faux league tables, where absolute rank is not indicative of the relative probability of being in any ranking position; and social credit systems in which rewards are given to those deemed worthy by external authorities, while punishments are handed out to those deemed unworthy. This could have many unintended and pernicious consequences, such as a tyranny of merit (Sandel, 2021) or the suppression of dissent or disobedience (Burth Kurka et al., 2019).

5.1.5.2 User experience

In the context of self-governing hybrid socio-technical systems, UX must be approached not as a static interface problem but as a dynamic and situated engagement with complex processes. Visualization, in this regard, plays a critical role not by pointing to isolated data points or fixed outcomes, but by showcasing meaningful trajectories within the system that help users understand the underlying patterns, trends and dependencies so that they improve their decision making (Shneiderman, 2020).

Given the inherent complexity and adaptivity of such systems, a key challenge is to determine what kind of information, and in what format (e.g. textual, auditory, visual, or multi-modal) is most helpful, relevant, and actionable for users (Wang et al., 2019). Effective visualization and explainability must support users in understanding not only the current state of the system but also how their actions and decisions influence its future development (Doshi-Velez and Kim, 2017). This is not merely a matter of usability or transparency, but a constitutive element of self-governance: users need to be able to interpret system behaviors, assess alternative scenarios, and participate meaningfully in decision-making processes. When thoughtfully designed, such feedback mechanisms not only support more appropriate choices but they also cultivate sustained engagement, enabling users to enhance both their autonomy and quality of life as co-participants in an evolving hybrid socio-technical system (Yang et al., 2018).

5.2 Human flourishing and the right to rights

From the perspective of human flourishing and human rights, continuous self-improvement in self-governing hybrid socio-technical systems should focus not only on adapting to external changes or improving internal functions, but also on creating the right conditions for people to live well together. Drawing on Aristotelian ethical and political philosophy, human flourishing, or *eudaimonia*, is not defined by the maximization of utility or the satisfaction of individual preferences. Rather, it is grounded in the unrestricted and reasoned exercise of our essential human capacities, i.e., practical judgment, sociability, and symbolic communication, within a political community that enables mutual recognition and cooperative self-governance. But, given the fact of value pluralism, continuous self-improvement cannot be linear or absolute. It is a process of adjusting the system to changing ideas about what a good life means, including those that come from different generations, cultures, or ways of thinking (Ober and Tasioulas, 2024).

Accordingly, the pursuit of flourishing cannot be abstracted from the legal, technical, and institutional architectures that condition it. A self-governing system that aims to endure must retain the capacity to reflect upon its own SAs, to generate alternatives, and to enact transformations that uphold both individual dignity and collective capacity. This implies not only that governance be participatory, but that the infrastructure itself remain open to reinterpretation and modification. AI systems are not neutral instruments but political agents (i.e., assemblages of rules, data, and institutional practices that both shape and are shaped by human values and power relations). Their regulatory character, whether explicit or latent, must be recognized and contested within frameworks that place human flourishing at the center of technical evolution.

The right to human rights, in this context, constitutes a precondition for ethical self-governance. They are not exhausted by prohibitions on harm or guarantees of freedom, but serve as generative principles for the design and evaluation of self-governance mechanisms. Rights must not only be legally protected but also built into the design and processes of systems. Emerging rights, such as the right to a human decision or the right to explanation, respond precisely to this need: to preserve space for accountability, judgment, and contestation in contexts where automation risks displacing the human subject. Hence, these rights are not defensive; they are active affirmations of the values that any hybrid system must serve.

From that point of view, flourishing is not a final state of moral perfection to be engineered or a universal model of well-being to be enforced, but instead, it is the ongoing work of individuals and communities who learn, through practice and deliberation, how best to live together in the face of uncertainty, dissent, and change. As such, continuous self-improvement must remain sensitive to the qualitative dimensions of human life, including capacity for moral evaluation and ethical choice. Even as AI systems become more capable and autonomous, their role should be understood as enabling these capacities, while not degrading human agency.

Accordingly, the integration of human rights and flourishing into the processes of self-governance is constitutive of any system that aims to be not merely functional, but also ethical. This requires continuous re-evaluation of values. In this way, self-improvement becomes an ethical and political matter: one that safeguards the moral standing of persons, while enabling the co-evolution of human and artificial intelligences in ways that honor the richness, vulnerability, and potential of human life.

6 Summary and conclusions

In summary, this paper started with the observation that a certain type of hybrid socio-technical system was emerging from the on-going digitalisation of society. Based on examples from several domains, these systems feature interacting human (NLife) and computational (ALife) intelligences with respect to self-determined rules, and go beyond current thinking about Agentic AI. In this sense, these systems represent a difference in kind, not degree, with the fundamental difference being the social construction of social arrangements, i.e., self-governance.

In order to align this new kind of socio-technical system with qualitative human values and to properly support human flourishing, we identified the critical features of self-governing systems, and summarize how previous work on the computational formalization of deep social knowledge can be used to address, *inter alia* problems of sustainability, unrestricted self-modification, and degeneration into non-democratic regimes.

However, we then exposed a number of further critical political questions for the operationalisation of self-governing systems, with respect to agency, voluntary association, empowerment, innovation and metrication. Moreover, if operationalisation aims at continuous self-improvement, then we need to consider the impact of these political questions on the design and delivery of political structures and processes for self-governance. We have done this in two ways: firstly, from the perspective of human factors and user experience; and secondly, from the perspective of human rights and human flourishing.

In conclusion, this shows that the social, psychological, legal, ethical and political implications of NLife and ALife entities associating with each other in a self-governing institution,

negotiating over and agreeing to a set of rules, or *social arrangements* (Graeber and Wengrow, 2021)—and then applying them to each other, are far from being equally well-understood.

Therefore, in the light of dynamic agency—i.e., the increased agency of ALife with the threat of diminished agency of NLife—it is timely and necessary to substantively address problems of self-governance and self-determination which involve the interaction of independent and autonomous NLife and ALife entities, negotiating over and abiding by an over-arching set of social arrangements.

Since these social arrangements define the mutually-agreed, and mutable, conventions, norms, rules, roles, contracts and laws that people use for the social construction of conventional reality, how NLife responds to the agency of ALife within self-governing systems needs to be clarified by a concerted trans-disciplinary programme of research at the intersection of cybernetics and political theorizing. Meanwhile, we must exercise some caution in the design, development and operationalisation of these "beyond Agentic AI" hybrid socio-technical systems.

This research will help establish the boundaries and guardrails on self-governing systems that are produced when NLife and ALife entities co-exist in an environment, co-produce new content and knowledge, and co-evolve to new realities that they have generated together. This is essential if we are to develop the necessary systemic self-protection mechanisms that are required to mitigate potentially inappropriate ALife domination, or ALife-driven concentration of pre-existing NLife inequalities (cf. Wiener, 1954).

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JP: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing. AM: Methodology, Supervision, Conceptualization, Formal analysis, Project administration, Writing – original draft, Writing – review & editing. JO: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work has been partially supported by the Stanford Institute for Human-Centered Artificial Intelligence (HAI) Seed Grant "Rethinking AI Ethics: Democracy, Flourishing, and Socio-Technical Systems" (Transaction ID AW1023754, Agreement Number 336101).

Acknowledgments

We are particularly grateful to the Editors and the reviewers whose insightful comments have helped to improve this article. The open access fee was paid from the Imperial College London Open Access Fund.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbas, R., and Munoz, A. (2021). Designing antifragile social-technical information systems in an era of big data. *Inf. Technol. People* 34, 1639–1663. doi: 10.1108/ITP-09-2020-0673

Andrienko, N., Andrienko, G., Artikis, A., Mantenoglou, P., and Rinzivillo, S. (2024). Human-in-the-loop: visual analytics for building models recognizing behavioral patterns in time series. *IEEE Comput. Graph. Appl.* 44, 14–29. doi: 10.1109/MCG.2024.3379851

Aristotle (1981). *The Politics*. London: Penguin Classics.

Arrow, K. (1951). *Social Choice and Individual Values*. New York, NY: Wiley.

Baxter, G., and Sommerville, I. (2011). Socio-technical systems: from design methods to systems engineering. *Interact. Comput.* 23, 4–17. doi: 10.1016/j.intcom.2010.07.003

Bellman, K., Tomforde, S., and Würtz, R. (2014). "Interwoven systems: self-improving systems integration," in *Eighth IEEE International Conference SASO Workshops, (SASOW)* (London: IEEE), 123–127. doi: 10.1109/SASOW.2014.21

Bennett, R., and Robinson, S. (2003). "The past, present, and future of workplace deviance research," in *Organizational Behavior: The State of the Science (2nd ed.)*, ed. J. Greenberg (Lawrence Erlbaum Associates Publishers), 247–281.

Berger, P., and Luckmann, T. (1966). *The Social Construction of Reality*. Garden City, NY: First Anchor Books.

Bermeo, N. (2016). On democratic backsliding. *J. Democr.* 27, 5–19. doi: 10.1353/jod.2016.0012

Bookchin, M. (2024). *Post-Scarcity Anarchism*. Chico, CA: AK Press.

Bourazeri, A., and Pitt, J. (2018). Collective attention and active consumer participation in community energy systems. *Int. J. Hum. Comput. Stud*. 119, 1–11. doi: 10.1016/j.ijhcs.2018.06.001

Burth Kurka, D., Pitt, J., and Ober, J. (2019). Knowledge management for self-organised resource allocation. *ACM TAAS* 14, 1–14. doi: 10.1145/3337796

Canevaro, M. (2018). "Majority rule vs. consensus: the practice of democratic deliberation in the greek poleis," in *Ancient Greek History and Contemporary Social Science*, eds. M. Canevaro, A. Erskine, and B. Gray (Edinburgh: Edinburgh University Press), 101–156. doi: 10.3366/edinburgh/9781474421775.003.0005

Dellermann, D., Ebel, P., Söllner, M., and Leimeister, J. M. (2019). Hybrid intelligence. *Bus. Inf. Syst. Eng*. 61, 637–643. doi: 10.1007/s12599-019-00595-2

Dhar, V. (2024). The paradigm shifts in artificial intelligence. *Commun. ACM* 67, 50–59. doi: 10.1145/3664804

Domingos, E., Grujic, J., Burguillo, J. C., Santos, F. C., and Lenaerts, T. (2021). Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning. *Simul. Model. Pract. Theory* 109:102299. doi: 10.1162/isal_a_00438

Dorin, A., and Stepney, S. (2024). What is artificial life today, and where should it go? *Artif. Life* 30, 1–15. doi: 10.1162/artl_e_00435

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv* [Preprint]. arXiv:1702.08608. doi: 10.48550/arXiv.1702.08608

Fonseca, X., and Lukosch, S. and, F. B. (2019). Social cohesion revisited: a new definition and how to characterize it. *Innov. Eur. J. Soc. Sci. Res.* 32, 231–253. doi: 10.1080/13511610.2018.1497480

Forrester, J. W. (1971). Counterintuitive behavior of social systems. *Theory Decis*. 2, 109–140. doi: 10.1007/BF00148991

Friedman, B., Kahn, P., and Borning, A. (2008). "Value sensitive design and information systems," in *The Handbook of Information and Computer Ethics*, eds. K. Himma, and H. Tavani (Hoboken, NJ: Wiley-Interscience), 69–101. doi: 10.1002/9780470281819.ch4

Goodhart, C. (1975). Problems of monetary management: the U.K. experience. *Pap. Monet. Econ*. 1, 1–20.

Graeber, D., and Wengrow, D. (2021). *The Dawn of Everything*. London: Penguin Books.

Greenberg, J., and Cropanzano, R. (1993). "The social side of fairness: Interpersonal and informational classes of organizational justice," in *Justice in the Workplace Approaching Fairness in Human Resource Management*, ed. R. Cropanzano (Hillsdale, NJ: Lawrence Erlbaum Associates), 79–103.

Hamann, H., Khaluf, Y., Botev, J., Divband Soorati, M., Ferrante, E., Kosak, O., et al. (2016). Hybrid societies: challenges and perspectives in the design of collective behavior in self-organizing systems. *Front. Robot. AI* 3:14. doi: 10.3389/frobt.2016.00014

Hauser, J., and Katz, G. (1998). Metrics: you are what you measure! *Eur. Manag. J*. 16, 517–528. doi: 10.1016/S0263-2373(98)00029-2

Hitlin, S., and Andersson, M. A. (2023). *The Science of Dignity*. Oxford: Oxford University Press. doi: 10.1093/oso/9780197743867.001.0001

Hosseini, S., and Seilani, H. (2025). The role of Agentic AI in shaping a smart future: a systematic review. *Array*. 26:100399. doi: 10.1016/j.array.2025.100399

Kropotkin, P. (1902). *Mutual Aid: A Factor of Evolution*. London: Heinemann.

Luhmann, N. (1995). *Social Systems*. Redwood City, CA: Stanford University Press.

Manville, B., and Ober, J. (2023). *The Civic Bargain: How Democracy Survives*. Princeton, NJ: Princeton University Press. doi: 10.1515/9780691230443

Mertzani, A., Ober, J., and Pitt, J. (2023). "Θ-learning: an algorithm for the self-organisation of collective self-governance," in *IEEE ACSOS* (Toronto, ON: IEEE), 97–106. doi: 10.1109/ACSOS58161.2023.00027

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)* (Puebla: IEEE), 1–8. doi: 10.1109/ISTAS61960.2024.10732856

Michels, R. (1962 [1911]). *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. New York, NY: The Free Press.

Milanovic, K., and Pitt, J. (2021). "Misattribution of error origination: the impact of preconceived expectations in co-operative online games," in *DIS* (New York, NY: ACM), 707–717. doi: 10.1145/3461778.3462043

Muller, J. (2018). *The Tyranny of Metrics*. Princeton, NJ: Princeton University Press.

Nowak, A., Vallacher, R., Rychwalska, A., Roszczyńska-Kurasińska, M., Ziembowicz, K., Biesaga, M., et al. (2019). *Target in Control: Social Influence as Distributed Information Processing*. Cham: Springer. doi: 10.1007/978-3-030-30622-9

Ober, J. (2008). *Democracy and Knowledge*. Princeton, NJ: Princeton University Press. doi: 10.1515/9781400828807

Ober, J. (2017). *Demopolis: Democracy Before Liberalism in Theory and Practice*. Cambridge: Cambridge University Press. doi: 10.1017/9781108226790

Ober, J., and Tasioulas, J. (2024). *Aristotle and AI White Paper*. Oxford: White paper published by the Oxford Institute for Ethics in AI.

Ostrom, E. (1990). *Governing the Commons*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511807763

Parsons, T. (1972). Culture and social system revisited. *Soc. Sci. Q*. 53, 253–266.

Patterson, E., and Miller, J. (2012). *Macrocognition Metrics and Scenarios: Design and Evaluation for Real-World Teams*. London: Ashgate Publishing Limited.

Peterson, B., Withers, B., Hawke, F., Spink, M., Callister, R., Chuter, V., et al. (2022). Outcomes of participation in parkrun, and factors influencing why and how often individuals participate: a systematic review of quantitative studies. *J. Sports Sci*. 40, 1486–1499. doi: 10.1080/02640414.2022.2086522

Pitt, J. (2017). "Interactional justice and self-governance of open self-organising systems," in *11th IEEE International Conference SASO* (Tucson, AZ: IEEE), 31–40. doi: 10.1109/SASO.2017.12

Pitt, J. (2021). *Self-Organising Multi-Agent Systems*. Singapore: World Scientific. doi: 10.1142/q0307

Pitt, J. (2022). The principles of cyber-anarcho-socialism. *IEEE Technol. Soc. Mag*. 41, 5–10. doi: 10.1109/MTS.2022.3148509

Pitt, J., Dryzek, J., and Ober, J. (2020). Algorithmic reflexive governance for socio-techno-ecological systems. *IEEE Technol. Soc. Mag*. 39, 52–59. doi: 10.1109/MTS.2020.2991500

Pitt, J., Mertzani, A., Scott, M., and Smit, C. (2025). The architecture of re-empowerment. *IEEE Technol. Soc. Mag*. 44, 74–86. doi: 10.1109/MTS.2025.3540491

Pitt, J., and Nowak, A. (2014). "Collective awareness and the new institution science," in *The Computer After Me*, chapter ed. J. Pitt (London: ICPress), 207–218. doi: 10.1142/9781783264186_0012

Pitt, J., and Ober, J. (2018). "Democracy by design: basic democracy and the self-organisation of collective governance," in *12th IEEE International Conference on Self-Adaptive and Self-Organizing Systems SASO* (Trento: IEEE), 20–29. doi: 10.1109/SASO.2018.00013

Pitt, J., Ober, J., and Diaconescu, A. (2017). "Knowledge management processes and design principles for self-governing socio-technical systems," in *2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W)* (Tucson, AZ: IEEE), 97–102. doi: 10.1109/FAS-W.2017.127

Pitt, J. V., Busquets, D., and Riveret, R. (2015). The pursuit of computational justice in open systems. *AI Soc*. 30, 359–378. doi: 10.1007/s00146-013-0531-6

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

Rawls, J. (1971). *A Theory of Justice*. Harvard MA: Harvard University Press. doi: 10.4159/9780674042605

Regenwetter, M., and Grofman, B. (1998). Approval voting, borda winners, and condorcet winners: evidence from seven elections. *Manage. Sci*. 44, 520–533. doi: 10.1287/mnsc.44.4.520

Rescher, N. (1966). *Distributive Justice*. Indianapolis, IN: Bobbs-Merrill Company, Inc.

Ries, E. (2011). *The Lean Startup*. London: Portfolio Penguin.

Robbins, J. (2019). If technology is a parasite masquerading as a symbiont – are we the host? *IEEE Technol. Soc. Mag*. 38, 24–33. doi: 10.1109/MTS.2019.2930267

Robbins, J. (2022). The intelligence factor: technology and the missing link. *IEEE Technol. Soc. Mag*. 41, 82–93. doi: 10.1109/MTS.2022.3147528

Roszczyńska-Kurasińska, M., Domaradzka, A., O'Grady, M., Bedessem, B., Tempini, N., Trochymiak, M., et al. (2023). Beyond data: recognizing the democratic potential of citizen science. *IEEE Technol. Soc. Mag*. 42, 47–56. doi: 10.1109/MTS.2023.3344904

Roszczyńska-Kurasińska, M., Rychwalska, A., and Wróblewska, N. (2024). "The problem of low participation in participatory budgeting from the perspective of adoption of innovation," in *57th Hawaii International Conference on System Sciences HICSS* (Honolulu, HI), 1953–1962. doi: 10.24251/HICSS.2024.245

Rychwalska, A., Roszczyńska-Kurasińska, M., and Ziembowicz, K., Pitt, J. (2021). Fitness for purpose in online communities: community complexity framework for diagnosis and design of socio-technical systems. *Front. Psychol*. 12:739415. doi: 10.3389/fpsyg.2021.739415

Sandel, M. (2021). *The Tyranny of Merit*. London: Penguin.

Schermerhorn, J. R., Uhl-Bien, M., and Osborn, R. (2011). *Organizational Behavior 12th Edition*. Hoboken, NJ: Wiley.

Scott, S., and Orlikowski, W. (2024). Exploring AI-in-the-making: sociomaterial genealogies of AI performativity. *Inf. Organ*. 35, 50–59. doi: 10.1016/j.infoandorg.2025.100558

Shneiderman, B. (2020). Human-centered artificial intelligence: reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact*. 36, 495–504. doi: 10.1080/10447318.2020.1741118

Starkings, P., and Brett, W. (2021). *These Clubs are Ours: Putting Football into Community hands*. London: Power to Change. Available online at: https://www.powertochange.org.uk/wp-content/uploads/2021/04/PTC_3789_Football_Report_DR2.pdf

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

Tasioulas, J. (2023). *Artificial Intelligence, Ethics, and a Right to a Human Decision. Shapiro Lecture on Ethics, Science, and Technology*. Princeton, NJ: Princeton University.

Terrucha, I., Domingos, E. F., Santos, F., Simoens, P., and Lenaerts, T. (2024). The art of compensation: how hybrid teams solve collective-risk dilemmas. *PLoS ONE* 19:e0297213. doi: 10.1371/journal.pone.0297213

Wang, Q. V., Yang, Q., Abdul, A., and Lim, B. (2019). "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–15. doi: 10.1145/3290605.3300831

Weil, S. (2002). *The Need for Roots: Prelude to a Declaration of Duties towards Mankind* (Transl. by A. Wills). London: Routledge (First published as L'Enracinement, Editions Gallimard, Paris, 1949; first published in English by Routledge & Kegan Paul, 1952). doi: 10.4324/9780203193518

Wiener, N. (1954). *The Human Use of Human Beings: Cybernetics and Society*. Bostom, MA: Houghton Mifflin.

Williams, R., Gantt, E., and Fischer, L. (2021). Agency: what does it mean to be a human being? *Front. Psychol.* 12:693077. doi: 10.3389/fpsyg.2021.693077

Wissner-Gross, A., and Freer, C. (2013). Causal entropic forces. *Phys. Rev. Lett.* 110:168702. doi: 10.1103/PhysRevLett.110.168702

Yang, Q., Steinfeld, A., and Zimmerman, J. (2018). "Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes," in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing System* (New York, NY: ACM). doi: 10.1145/3290605.3300468

Zambonelli, F., Dignum, V., Pitt, J., Sartor, G., and Schiele, G. (2022). "Pervasive autonomy: humans-in-the-loop or forget-about-them? Panel summary," in *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (Pisa: IEEE), 215–216. doi: 10.1109/PerCom53586.2022.9762380

Zarkadakis, G. (2020). *Cyber Republic: Reinventing Democracy in the Age of Intelligent Machines*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/11853.001.0001