Algorithmic Reflexive Governance for Socio-Techno-Ecological Systems

Jeremy Pitt, Dept. of Electrical & Electronic Eng., Imperial College London, UK
John Dryzek, Institute for Governance and Policy Analysis, University of Canberra, Australia
Josiah Ober, Dept. of Political Science, Stanford University, CA, USA

Abstract. This position papers develops an algorithmic approach to deliberation as the basis for institutional reflexivity. We propose the idea of *algorithmic reflexive governance* for *socio-techno-ecological systems*, bringing together inter-disciplinary ideas from Artificial Intelligence and Democratic Theory in systems that recognize the interactions between social, technological and ecological actors and processes. We argue that this innovative approach to conservation and sustainability can make a significant contribution to mitigating the climate emergency.

Introduction

The dominant social, economic and political institutions of the early 21st Century seem to have difficulty recognising environmental preconditions and impacts, and are either insensitive to the trajectory towards catastrophe or lack the ability to adapt in a timely and positive fashion. One diagnosis for this "institutional sclerosis" is a lack of reflexivity [DryPic].

Reflexivity, generally, is the ability of a structure, process, or organisation to reconfigure itself in response to reflection upon its own performance. The core aim of this work is to develop self-governing socio-technical systems where reflexivity with respect to ecological and environmental impact is enabled through algorithms for deliberation, introspection and self-organisation. To this end, we begin with a discussion of the four dimensions of reflexivity, proceed to discuss an algorithmic approach to deliberate processes which resolve tensions in each of the four dimensions, and finally use this as the recommended basis for developing socio-techno-ecological systems.

Four Dimensions of Reflexivity

In the context of environmental governance, it has been argued [DryPic] that there are four key dimensions of reflexivity, each of which pits two plausible drivers in tension (see Figure 1):

- Sources of knowledge: public participation vs. expertise;
- Composition of public discourse: diversity vs. consensus;
- Institutional architecture: polycentricity vs. centralization; and
- Institutional dynamics: flexibility vs. stability.

It is further argued [DryPic] that deliberative processes can handle the tensions between each of these binaries. In particular, deliberation can contribute to reflexive governance by managing these tensions respectively: by joining expert and lay deliberation, either face-to-face or by integrating the two in a broader deliberative framework (e.g. citizen assemblies based on sortition to provide political balance and inclusivity); by generating meta-consensus, through validation of multiple perspectives while establishing and maintaining practicable agreements between them; by enabling deliberative learning through sharing information, experience or best-practice, and by transcending the binary through meta-deliberation, i.e. considering the effectiveness of the deliberative properties of the system; and by requiring periodic deliberative scrutiny of any stable institutional arrangements.

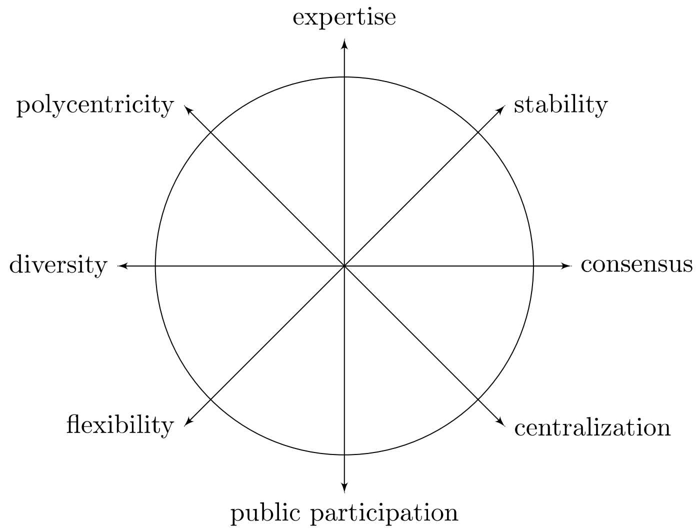

Figure 1: *The Four Dimensions of Reflexivity*

This work proposes that these processes can be specified as algorithms.

Algorithmic Reflexive Governance

The idea behind algorithmic reflexive governance is to specify and implement algorithms for each of the deliberative processes defining the four dimensions of reflexivity. This specification of deliberative processes as algorithms leverages several prior works from self-organising systems, including: studies of *relevant expertise aggregation* [Ober13] to join expert and lay decision-making; *interactional justice* [Pitt17] as a way of achieving meta-consensus; situating a system-of-systems within the *zone of dignity* [Ober17] as a form of meta-deliberation, and the use of *quasi-stability* [Rawls] to scrutinize and (if necessary) restore stability. We examine each of these processes in turn.

Relevant Expertise Aggregation (REA)

In the first dimension, relevant expertise aggregation (REA) is proposed in order to address the tension between majority preference and expert judgement [Ober13]. REA is a process of decision-making that is both ‘democratic’ (in the sense that it takes into account the expressed preferences of the majority) and ‘epistemic’ (in the sense that it takes into account the knowledge and judgement of experts). The deliberative process is that expert groups make recommendations and citizen groups select policy priorities based on those recommendations.

A procedure is informally described in [Ober13]. A governing institution is faced with a set of problems that need to be addressed. For each problem, it identifies a limited number of domains: each domain is ranked in [0,1] and weighted (with the sum of weights on the domains normalised to 1) according to its perceived relevance to the issue. A group of experts are consulted: it is supposed that these experts are reliable, i.e. they have good reasons to disclose their true opinion, and have a probability better than random of reaching the ‘correct’ decision (taking advantage of the Condorcet Jury Theorem [Condorcet]). The group of domain experts is offered a limited number of policy options, each option being understood as a potential solution to the issue. Each option is given a final score as a proportion of the expert votes multiplied by the weight of the domain, and summed over all domains. The score is used to determine policy selection, which can be done a number of different ways, depending on political regime.

A formal algorithmic specification, and its implementation using multi-agent systems, was used to select policies for regulating community energy systems [REA] (e.g. given uncertainty in renewable energy generation, how much energy to import, export, generate, store, etc). In this simulation, there are a set of criteria for evaluating policies, and for each criterion there is a set of metrics is evaluating it. Each agent 'knows' a subset of criteria, and a subset of the related metrics. At regular intervals, a number of 'expert' groups, comprising a subset of the agents, were formed. Given a starting policy (which could either be an initial policy or some random policy), each group of agents used machine learning algorithms to produce a recommendation for the 'best' policy option under the existing and expected operational conditions. The collective agents then voted for the system policy for next operating interval.

Interactional Justice

For the second dimension of reflexivity, interactional justice might address the tension between diversity and consensus. Interactional justice in sociology [SUO] and organizational theory [GC] deals with the interpersonal treatment of the members of an institution according to the outcomes produced by its rules, structures and decision-makers. Therefore, interactional justice is a user-centric aspect of justice that can be used for realising values of fairness and inclusivity in organisations and communities. It has been shown how the concept of interactional justice can be applied to self-organising electronic institutions composed of computational components (i.e. software agents), where it is aimed at realising shared *values* amongst a set of disparate agents who are members of the same (electronic) institution [Pitt17].

This use of interactional justice in self-organising electronic institutions was composed of three phases of interaction in the context of institutional rules. In the first phase, a subjective evaluation was made of the interpersonal treatment of the individual vis-a-vis its values and the outcomes of the institution's decisions, i.e. each agent computed a subjective assessment of the 'quality' of an institution with respect to personal values. In the second phase, these individual assessments were pooled to form a collective assessment of the institution's quality with respect to the collective's shared values (and the priorities on them); a social network enabled communication between individuals to disseminate and compare personal treatments by the institution with respect to these shared values. The third phases addresses how this group of agents react if they collectively think that their treatment has been 'unfair': in particular, reformation of the institution if the collective experience was considered incongruent with the operational policies.

The specification and implementation of an algorithm for this form of interactional justice in a multi-agent system used ideas from *distributive justice* [Rescher] for the personal self-assessment; and then used *opinion formation over a social network* [RamCanoPitt] to aggregate subjective (diverse) self-assessments of fairness into a collective (consensual) assessment, as an indicator of the institutional quality. Each agent had certain values, specifically 'fairness', and a set of metrics that it used to evaluate whether or the outcome of institutional deliberation was 'fair' (the specific application was resource allocation, so the metrics included legitimate claims, utility, and satisfaction). From this and interaction with its neighbours, each agent formed an initial opinion on fairness. These opinions then percolated through the network, and agents updated their opinions accordingly. Their final opinions could be used to motivate institutional change, if the collective opinion was one of specifically unfairness, or more generally, an incongruence between institutional rules and experience. Linking back to relevant expertise aggregation, one option would be a majority vote on a policy to address the problem, but another could be some form of 'civil' disobedience as a mechanism of bringing about positive change [Kurka] (and see below).

The Zone of Dignity

In a system of systems, the relationship between systems (i.e. the institutional architecture) can (in extremis) be either hierarchical (single centralized decision-maker) or polycentric (multiple autonomous centres of decision-making). In a hierarchical system, loosely speaking, information flows 'up' and policies flow 'down'. In an ideal world, perfect information drives appropriate policy formation, i.e., evidence-based policy-making. In a sub-ideal world, information is abstracted up the hierarchy, and policies are reinterpreted down the hierarchy.[1] To counter this, some forms of corporate governance have introduced the concept of *guardrails*. The idea of guardrails is that in a hierarchy, the 'upper' (management) layer specifies that within specific boundaries, the 'lower' (implementation) layer can make any decision, or take any action, they want, in order to complete a given task. Moreover, the guardrails are not necessarily fixed: if the 'lower' layer needs to go outside the guardrails, it seeks permission from the 'upper' layer.

The idea of a guardrail can be seen within Ostrom's Institutional Design Principle 7, which states that, in a system of systems, the *minimal recognition of the right to self-organise* should be observed [Ostrom]. In fact, many of the failures to sustain a common-pool resource, reported in [Ostrom], could be traced to a failure of this principle. The idea of a guardrail is also present in Ober's Demopolis [Ober17], in a discussion of the continuum of distributive justice from full equality to complete liberty. Dignitarian considerations set limits on how far the libertarian and egalitarian tendencies can push equality and liberty. The avoidance of indignity for a democratic regime defines the *zone of dignity* (ZoD), which determines the acceptable range of policy options for distributive justice, i.e., not too coercive, not too heartless.[2] The ZoD therefore represents a kind of preference meta-consensus (or guardrails) on the acceptable range of (institutional) policy options.

To specify and implement a deliberative process for algorithmic reflexive governance based on the ZoD, we propose three components: a *coordinate plane*, *metrics* to locate an object in that coordinate plane, and *meta-rules* to control the trajectory of an object (i.e. a set of rules) as it moves within the plane. The proposed coordinate plane extends the ZoD from a linear one to a two-dimensional plane, by also including an axis of 'resource extraction' or taxation, that the

---

[1] One example: in the UK, a policy of *restorative justice*, a process that tries to resolve disputes to the mutual satisfaction of all concerned parties through negotiation, was introduced by a regional UK police force [Stockdale]. The policy was interpreted differently by different groups: senior management understood the theoretical concepts, key values and nuanced differences between "instant restorative disposals" and "restorative conferencing"; middle management were focused on performance; and uniformed officers 'on the beat' applied the policy as frontline practitioners. The latter group found themselves conflicted between pressures to resolve disputes without involvement of the criminal justice system, but also having to increase the number of 'detections' (the number of cases resolved with a ticket, charge, caution, etc.).

[2] This raises the question of why there is not a *maximal* recognition of the right to self-organise, as this implies that in a system of systems, unlimited rights to self-organise do not affect the sustainability of a local common-pool resource. Maybe this is true, but it does not necessarily augur well for the political regimes 'at the edge', i.e. the leaf node of a hierarchy or regions on the periphery of a state, e.g. some 1970s UK academic departments, organized crime, warlordism, feudalism, etc. See for example [Weber].

'upper' institution imposes on the 'lower' (cf. [Acemoglu]).[3] As shown in Figure 2, this characterises four quadrants of indignity that can bring about systemic collapse: exploitation, brought about by excessive tax and excessive control (tyranny); stagnation, brought about by inadequate tax and excessive control (no resources for top-down policies, no freedom-of-manoeuvre for bottom-up initiatives); devaluation, brought about by excessive tax and inadequate control (a system providing benefits without responsibilities is destined to fail); and fragmentation, brought about by inadequate tax and inadequate control (e.g. a weak state without a monopoly on violence produces factionalism (e.g. warlordism)). This plane defines the space in which to locate systemic configurations and evaluate their trajectories, and identifies the acceptable boundaries in this space.

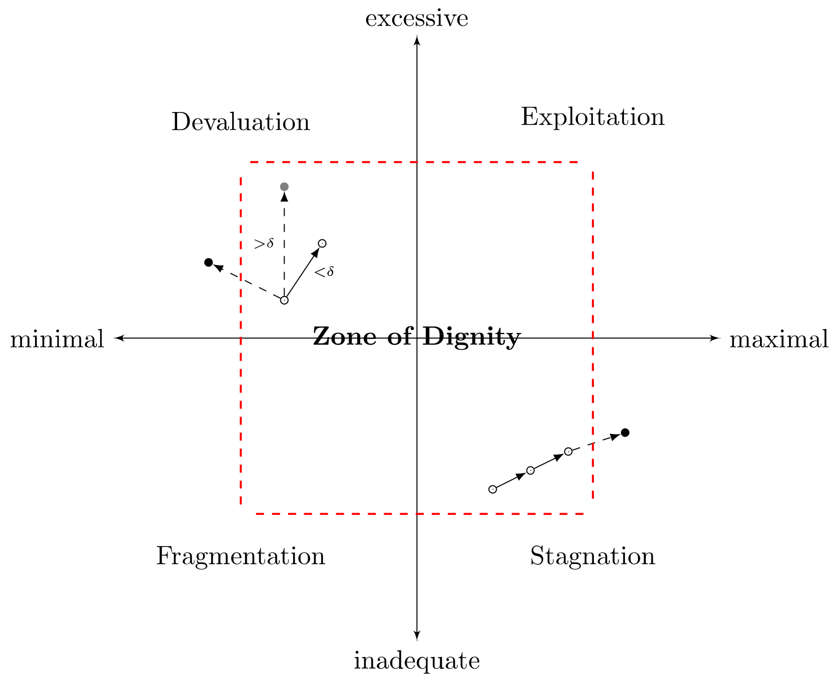

Figure 2: *The Zone of Dignity in Two Dimensions*

The second component is to use to apply ideas of procedural justice to the evaluation of institutional configurations to locate them within this space. Procedural justice is generally concerned with fairly, accurately and efficiently evaluating procedures and the actors enacting those procedures. It has been used, or example, in dispute resolution to determine the trade-off between 'adequate' participation in the process and the 'accuracy' of the outcome, and in public health, the determine the costs and benefits of the authorities imposing decisions on the populace. The procedural justice framework for evaluating self-organising electronic institutions proposed in [PBR] tried to metricate three principles: the *participation* principle, i.e. the purposeful activities in which actors take part in relation to governance (not just voting); the *transparency* principle, i.e., the amenability of procedures to be subject of investigation and analysis to establish facts of interest; and the *balancing* principle, i.e., the proportionality of relative benefits and burdens. For each principle, a number of metrics were proposed, for example, the participation principle was measured by empowerment, inclusivity, representation and consultation metrics.

---

[3] The notion of 'resource extraction' is intentionally general and covers the spectrum between "stationary bandits" [Olson] and meso-level water temples coordinating rice-planting [LK].

The third component is to apply ideas from metric spaces, as used in the specification of dynamic norm-governed multi-agent systems [Artikis]. The idea is to use the framework of procedural justice to identify an institutional configuration as a point in the ZoD, and then either to implement meta-rules on the movement in that plane, or to evaluate the trajectory of the configuration in that plane. These meta-rules could, for example, prohibit movement from one configuration to another if it exceeds a certain 'distance' (δ), or prohibit movement to a particular configuration altogether if it lies outside the boundaries of the ZoD (see upper left quadrant of Figure 2: dashed transitions are impermissible, either because the resulting configuration lies outside the ZoD, or the distance exceeds δ). Alternatively, observing the speed and direction of a configuration's trajectory might trigger reflective deliberation which alters its course before the anticipated trajectory leaves the ZoD (see lower right quadrant of Figure 2).[4]

Quasi-Stability

For institutional dynamics, we propose *quasi-stability* for controlling unrestricted self-modification. This notion of quasi-stability is derived from Rawls' theory of justice [Rawls], Ashby's design for a brain [Ashby], and Leibensteins's definition of economic equilibrium [Leibenstein]. Rawls asserted that "a well-ordered society is quasi-stable with respect to the justice of its institutions and the sense of justice needed to maintain this condition". His concepts of equilibrium and stability of a system were derived from Ashby, who defined, *inter alia*: polystable systems, as a system whose parts have many equilibria and random parts have been joined at random; ultrastable systems, as a particular type of reactive systems embedded in an environment; and multistable systems, a system whose state and behaviour increasingly depends on what happens to it, not the state in which it was initialized. From these definitions, the concept of quasi-stability was defined by Leibenstein as a system in which, after a period of disruption, some, but not all, of the control variables return to their prior state.

Algorithmics for institutional dynamics have been investigated, in part, in [Kurka, PO18]. Open systems with self-determined rules, whether social or electronic, have to deal with the expectation of error (e.g.: in a social system, excess appropriation from a common-pool resource; in an electronic system, interference in a wireless network), dissent, and potential incongruence. In [Kurka], these issues led to reorientation of either the rules or replacement of the rulers, depending on the nature of the error or incongruence (i.e. either the rule was wrong, or the ruler was applying the rule wrongly). In [PO18], incongruence led to a re-assignment of roles to ensure that all 'citizens' share the burden, in one way or another (e.g. equally), in the making, adjudicating and enforcing the rules. In both cases, some form of quasi-stability could be observed: the rules would be re-configured and the system would be stable for a period, with a different setting for the 'control variables', until another self-examination reset them again.

The key outcome of quasi-stability for ecological systems is that while a shift in social or environmental circumstances may cause an institution to become unjust – for example, a reaction to crisis may temporarily cause an institution to perceived as unjust, in due course the institution

[4] We note in passing that understanding the trajectory of a system is important to overcome the cognitive bias of "just noticeable difference" within a changing population. A series of "small enough" changes over a "long enough" period are effectively imperceptible to those that experienced them; while new entrants to the system think that "it was ever thus". However, the end state is far from the original state and much less preferable. The climate emergency and the centralization of academic administration are two examples of when the rate of change in a system synchronises with the rate of change of the population, its trajectory can approach and even leave the ZoD, terminating with systemic crisis and collapse.

is reformed as the crisis abates, and justice is restored. Therefore, it is a critical quality of a system that ensures reversion to *an* equilibrium, i.e. the (current) equilibrium is a product of what happened to the system, not the equilibrium from which it started.

Socio-Techno-Ecological Systems

We observe that while reflection requires human agency, ecosystemic reflexivity also requires the ability to listen to and interpret signals from the non-human (natural) world. The challenge for algorithmic reflexive governance is then two-fold. Firstly, there is the challenge to create *socio-techno-ecological* systems, by extending the non-human world with signal-generating sensors, and enriching the human world with intelligent software agents capable of both signal-processing and participation in the deliberative processes – but at certain levels of collective decision-making in social systems, "participation" at such levels means "supportive of", not "determinant of". Secondly, there is a challenge to demonstrate that the deliberative mechanisms/principles we advocate for resolution of the four tensions are not only consistent with each other, but also mutually reinforcing, such that these systems are sustainable.

Community energy systems (CES), which could be considered as an exemplar of socio-techno-ecological systems, illustrate the first challenge. A community energy system takes a primarily commons-oriented approach to energy generation, distribution, consumption and storage, rather than a market-oriented approach (cf. [Bollier]). The 'socio' component of the system comprises the human stakeholders; the 'ecological' component is concerned with ensuring that the actions of those stakeholders in accomplishing their systemic goals do not adversely affect their physical surroundings or other living organisms. In principle, SmartMeters should be the sensors that are inserted into the ecological system that are 'reading' the state of the environment, and should also be the software agents that support the human stakeholders in managing the balance between getting what they want out of the system without doing any environmental damage in the process. Artificial Intelligence, in the form of algorithmic deliberation and reasoning should be the techno 'glue' which links the socio- and the eco-logical.

However, four layers of interaction between human stakeholders and artificial intelligence have been identified in the management of energy as a common-pool resource (as, for example, in a CES). These are:

- *delegated* – humans out of the loop: some operations are controlled by artificial intelligence, which works without (or with limited) user awareness; for example, in CES, micro-grid load-balancing and voltage control can be implemented by regulating refrigerator compressor cycles;
- *programmable* – humans on the loop: daily operational requirements are specified by members of the CES, and artificial intelligence resolves the constraints (and disputes); for example, with programmable appliances such as dishwashers, the CES members specifies constraints (completion time, cost), and delegates to the appliances the task of negotiating a collectively agreeable schedule;
- *interactive* – humans in the loop: artificial intelligence detects problematic situations and indicates active intervention is required by humans; for example, a SmartMeter can initiate collective action for overload prevention, although any such intervention has to be sensitive to social situations (hence supportive not determiner); and
- *attentive* – humans *are* the loop: each stakeholder may occupy multiple different roles besides producer and consumer in a CES, for example, as an investor or a 'citizen', they will be called upon to consider investment decisions, engage in interactions with third parties, help resolve internal disputes, and so on.

Therefore, the deeper challenge for algorithmic reflexive governance in socio-techno-ecological systems is to manage the four binaries across all four layers. However, herein lies the roots of the second challenge. Of course, there are the practical aspects of implementing such a system; but there is a theoretical argument about the potential for success. Going back to Ashby's Design for a Brain [Ashby], it is argued that for stability (or in our case, sustainability) of a system, its regulator has to be more complex (have more states) than the system being regulated. Therefore, if by conjoining human and computational intelligence, we create a system that is more complex, and also more complex than the ecological system, it should be possible to regulate it; if, on the other hand, the joining of human and computational intelligence results in a diminution of human intelligence [Nowak, Robbins], then we are perhaps no better off, and possibly worse, if our humanity is diminished as well as depleting and despoiling the natural world.

Summary and Conclusions

In summary, we have proposed an algorithmic basis for implementing deliberative processes which can help to resolve the four tensions in reflexive governance. We have argued that algorithmic reflexive governance can monitor the impact of institutions on ecosystems, and vice versa, and is critical foundation for reshaping core values and reconfiguring current practices, especially when incongruence between those values and practices is exposed, especially when those values are concerned with environmental sustainability and communal well-being. In this way, we believe that algorithmic reflexive governance can sustain socio-techno-ecological system-of-systems within a Zone of Dignity that balances control (minimal/maximal recognition of the right to self-organise) against 'taxation' (minimal/maximal levels of resource-extraction for socially-productive coordination).

However, there remains much work to be done, and at least three advances need to be made. The first advance is integration of the four algorithmic processes within a common computational framework. The second advance is integration of this framework within a value-sensitive design framework, i.e. in the design of socio-techno-ecological systems, the values to be achieved and sustained by algorithmic reflexive governance should be considered as primary 'supra-functional' design requirements, not afterthoughts to be addressed after the system has been developed and deployed (cf. [PO18]). Thirdly, we need a demonstration of efficacy: an exemplar which, by bringing together ideas from Artificial Intelligence and Democratic Theory, convincingly demonstrates the sustainability of a socio-techno-ecological system.

Denial often follows three stages: firstly refusing to accept that there is a problem; secondly, acknowledging that there is a problem, but refusing to accept that its consequences will be "that bad"; and finally, conceding that there always was a problem and its consequences have been "that bad", but it is too late to do anything about it now. We refuse to accept that it is too late.

References

[DryPic] J. Dryzek and J. Pickering. Deliberation as a catalyst for reflexive environmental governance. Ecological Economics, 131:353–360, 2017.
[Ober13] J. Ober. Democracy's wisdom: An Aristotelian middle way for collective judgement, American Political Science Review, 107(1):104–122, 2013.
[Pitt17] J. Pitt. Interactional justice and self-governance of open self-organising systems. In 11th IEEE SASO, pp31–40, 2017.
[Ober17] J. Ober. Demopolis: Democracy before liberalism in theory and practice. Cambridge University Press, 2017.
[Rawls] J. Rawls, *A Theory of Justice*. Harvard University Press, 1971.

[REA] J. Pitt, R. Cardoso, E. Hart, and J. Ober. Relevant expertise aggregation for policy selection in collective adaptive systems. In 3rd FAS*W, pp136–141, 2018.
[Condorcet] C. List and R. Goodin. Epistemic Democracy: Generalizing the Condorcet Jury Theorem. Journal of Political Philosophy, 9(3):277-306, 2001.
[SUO] J. R. Schermerhorn, M. Uhl-Bien, and R. Osborn, *Organizational Behavior 12th Edition*. Wiley, 2011.
[GC] J. Greenberg and R. Cropanzano, "The social side of fairness: Inter- personal and informational classes of organizational justice," *Justice in the workplace*, 1993
[Rescher] N. Rescher, *Distributive Justice*. Bobbs-Merrill, 1966.
[RamCanoPitt] D. Ramirez-Cano and J. Pitt, Follow the leader: Profiling agents in an opinion formation model of dynamic confidence and individual mind-sets, *IAT*, pp. 660–667, 2006.
[Kurka] D. Burth Kurka, J. Pitt, P. Lewis, Al Patelli and A. Ekárt. Disobedience as a Mechanism of Change. IEEE SASO, pp1-10, 2018.
[Ostrom] E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.
[Acemoglu] D. Acemoglu. Politics and economics in weak and strong states. Journal of Monetary Economics, 52(7):1199-1226, 2005.
[Olson] M. Olson. Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships. Basic Books. 2000.
[LK] J. Lansing and J. Kremer. Emergent properties of Balinese water temple network: coadaptation on a rugged fitness landscape. American Anthropologist, 95:97–114, 1993.
[Stockdale] K. Stockdale. Police understandings of restorative justice—the impact of rank and role. Restorative Justice, 3(2):212-232, 2015.
[Weber] M. *Weber. Politics as a Vocation. Philadelphia: Fortress Press, 1965.*
[PBR] J. Pitt, D. Busquets, and R. Riveret, Procedural justice and 'fitness for purpose' of self-organising electronic institutions, in *Proc. PRIMA 2013*, ser. LNCS, vol. 8291. Springer, 2013, pp. 260–275.
[Artikis] A. Artikis, Dynamic specification of open agent systems, *Journal of Logic and Computation*, vol. 22, no. 6, pp. 1301–1334, 2012.
[Ashby] W. R. Ashby, *Design for a Brain*. Chapman-Hall, 1952.
[Leibenstein] H. Leibenstein. *Economic Backwardness and Economic Growth*. Wiley, 1957.
[PO18] J. Pitt and J. Ober. Democracy by Design: Basic Democracy and the Self-Organisation of Collective Governance. IEEE SASO, pp20-29, 2018.
[Bollier] D. Bollier and S. Helfrich. The Wealth of the Commons: A World Beyond Market and State. Levellers Press, 2012.
[Nowak] A. Nowak, P. Lukowicz and p. Horodecki. Assessing Artificial Intelligence for Humanity, IEEE Technology & Society Magazine, 37(4):26-34, 2018.
[Robbins] J. Robbins. If Technology Is a Parasite Masquerading as a Symbiont — Are We the Host? IEEE Technology & Society Magazine, 38(3):24-33, 2019.