Towards Socially Self-Aware Machines

Peter Lewis

p.lewis@aston.ac.uk

Let's talk about

Artificial Intelligence...

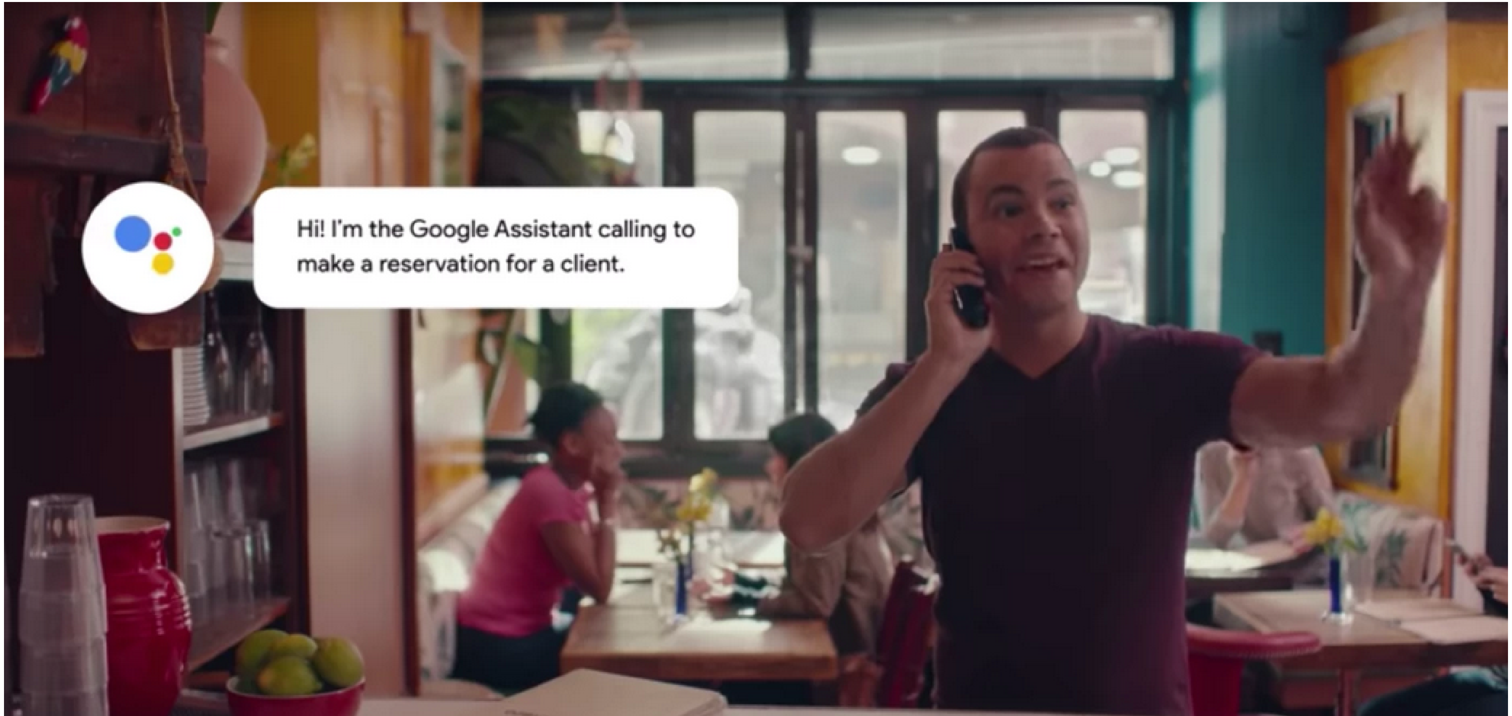
Yoshua Bengio

Yann LeCun

Geoffrey Hinton

AI-powered booking service Google Duplex rolls out to iOS & Android 5.0+ devices

Sarah Perez @sarahintampa / 2 weeks ago



AI-Powered, Self-Driving Robots Are Taking On a Bigger Role at Walmart Stores

The world's largest retailer is making a growing bet on robots and artificial intelligence to gain a competitive edge.

Danny Vena (TMFLifeIsGood)
Mar 19, 2019 at 7:31AM

Competition in the retail industry has never been more cutthroat. The dawn of e-commerce has caused a paradigm shift, with traditional retailers having to change with the times or fall by the wayside.

Walmart (NYSE:WMT) is representative not only of the old guard of retail, but also of the transition that is happening among brick-and-mortar stores to adapt to this new reality. In addition to a fierce move into e-commerce, the once-stodgy retailer has embraced cutting-edge technology to help keep costs in check and provide a better shopping experience for its customers.

Case in point: Self-propelled robots are now taking on an increasing role in Walmart's operations.

AUTHOR

Danny Vena (TMFLifeIsGood)

Daniel W. Vena, CPA, CGMA is long-term investor searching for intangibles that provide explosive growth opportunities in his investments. He served on active duty with the US Army and has a Bachelors degree in accounting.

Follow @dannyvena 774 followers

ARTICLE INFO

Mar 19, 2019 at 7:31AM

Technology and Telecom

Qualcomm's new AI-powered camera design platform could popularize the use of computer vision

George Paul Apr. 11, 2019, 9:12 AM

- **This is an excerpt from a story delivered exclusively to Business Insider Intelligence IoT Briefing subscribers.**
- **To receive the full story plus other insights each morning, click here.**
- **Visit BusinessInsider.com for more stories.**

Qualcomm released its next-generation AI-powered camera design platform using its latest purpose-built systems on a chip (SoCs) to better enable on-device machine learning and processing capabilities.

LATEST THINKING

Share Print Download

THE NEW, NEW NORMAL: EXPONENTIAL GROWTH POWERED BY AI

LEARN MORE

OVERVIEW

Share

AI has the potential to allow companies to not only do different things, but also to do things differently. It represents a step change in balancing growth, profitability, sustainability and trust.

With 46 percent of executives indicating that AI will drive increased revenue opportunities, getting AI right makes companies more valuable to their ecosystem and helps them maintain a competitive position in a world that will soon be powered by AI.

But to realize real and sustainable benefits, AI must evolve from being a hot new trend to a seamless enabler — woven into the fabric of the enterprise and working alongside and augmenting people.

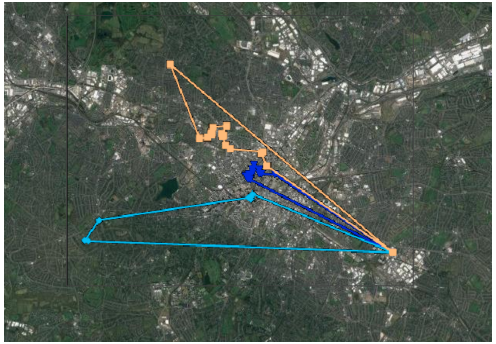
84%

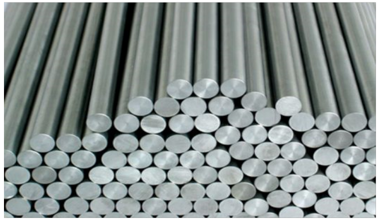
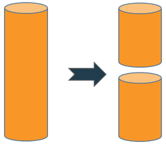
of organizations say AI will create new categories of products, services, business

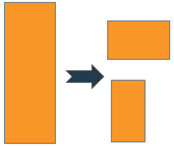
Aston University
BIRMINGHAM UK

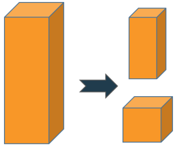
European Union
European Regional
Development Fund

Innovate UK

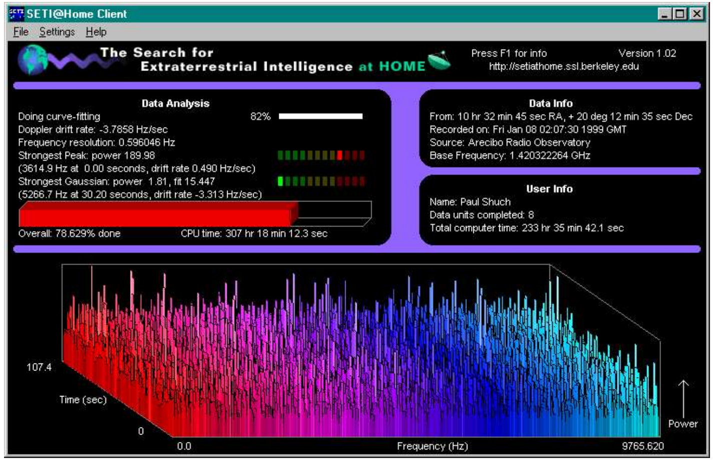



1D CUTTING

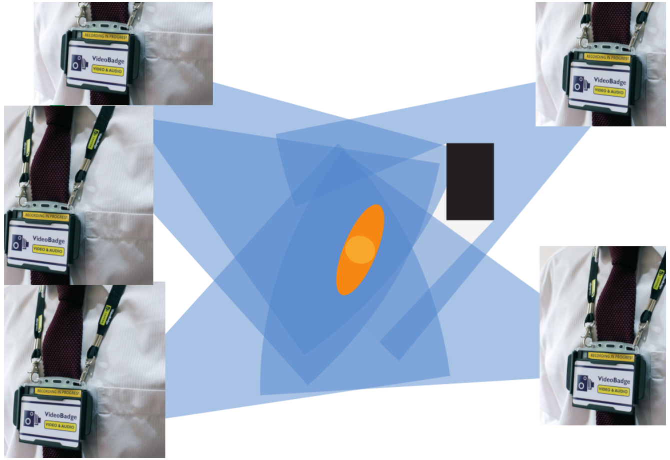


2D CUTTING

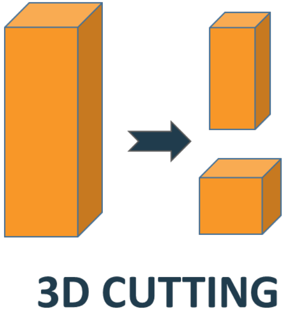


3D CUTTING

[1] Chitty DM, Wanner E, Parmar R, Lewis, PR. **Scaling ACO to large-scale vehicle fleet optimisation via Partial-ACO**. In: GECCO '19: Proceedings of the Genetic and Evolutionary Computation Conference Companion. ACM; 2019.
[2] Chitty DM, Parmar R, Lewis, PR. **Improving Urban Air Quality Through Long-Term Optimisation of Vehicle Fleets**. Advances in Intelligent Systems and Computing. Springer; 2019.

These Are Socio-Technical Systems

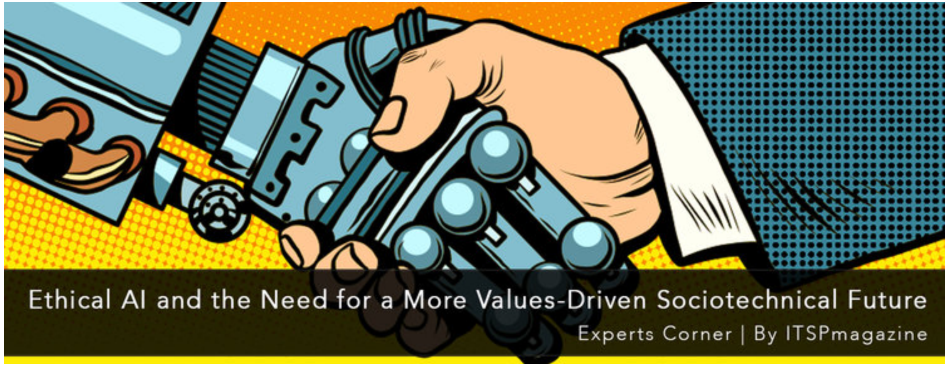






Intelligent Socio-Technical Systems?

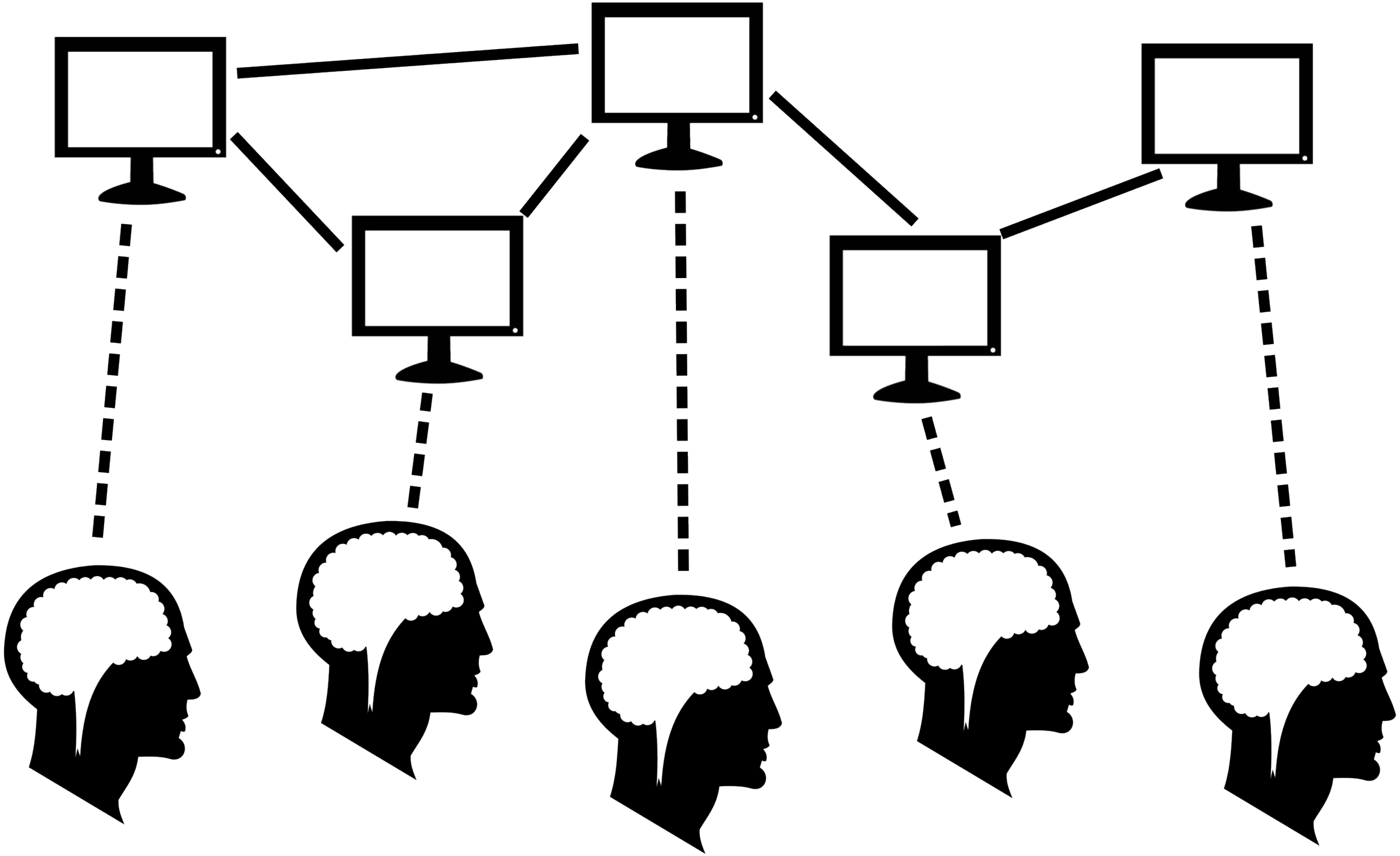
- These socio-technical systems provide **new opportunities for citizens to work together**, supported by machines, to tackle pressing societal challenges.
- As technology becomes ever more integrated into our daily lives, **social issues become at least as important as technical ones** in its design [3].
- From automation to autonomy $\longrightarrow$ **delegation of our autonomy** to 'intelligent' machines.



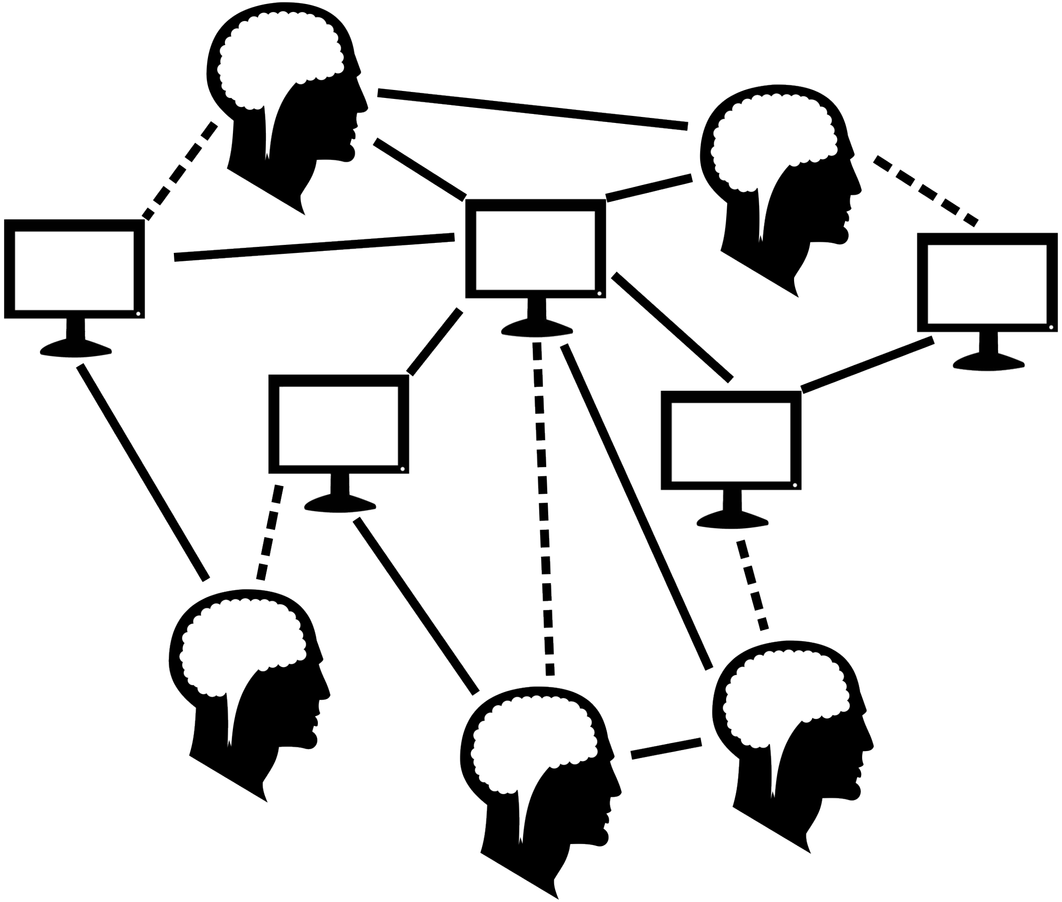
[3] Bellman KL, Botev J, Hildmann H, Lewis, PR, Marsh S, Pitt J, et al. **Socially-Sensitive Systems Design: Exploring Social Potential**. IEEE Technology and Society Magazine. 2017;36(3):72–80

Source: https://www.itspmagazine.com/from-the-newsroom/ethical-ai-and-the-need-for-a-more-values-driven-sociotechnical-future

Two Classes of Socio-Technical Systems



Two Classes of Socio-Technical Systems



Key Challenge

How to build autonomous systems that work well as part of society:

- **alongside us,**
- **and on our behalf?**

Autonomous systems that are:

- Able to manage competing priorities, while adapting to new unfolding situations,
- By-design congruent with social expectations, and their underlying human values, ethics, and norms [3].
- Foster and maintain appropriate forms of trust [4].

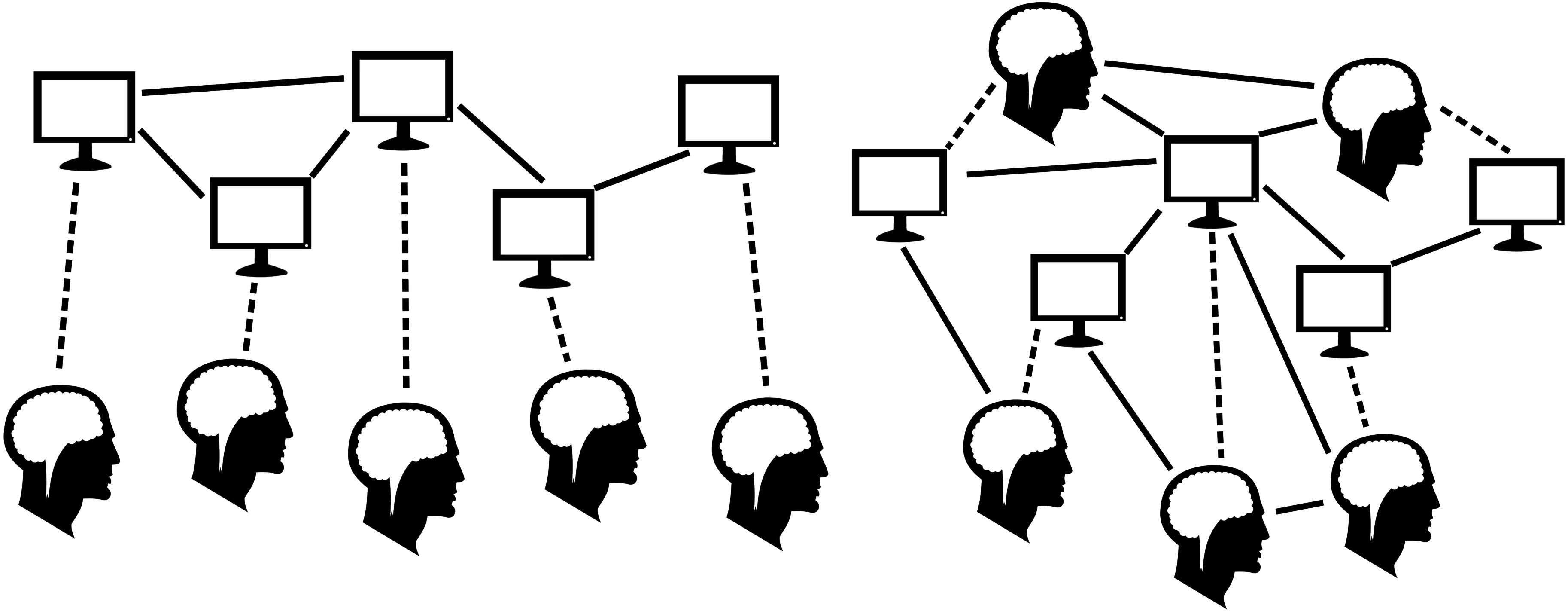
[3] Bellman KL, Botev J, Hildmann H, Lewis, PR, Marsh S, Pitt J, et al. **Socially-Sensitive Systems Design: Exploring Social Potential**. IEEE Technology and Society Magazine. 2017;36(3):72–80
[4] Andras P, Esterle L, Guckert M, Han TA, Peter R Lewis, Milanovic K, et al. Trusting Intelligent Machines. IEEE Technology and Society Magazine. 2018;37

Building Intuition: Outcomes We May Not Expect

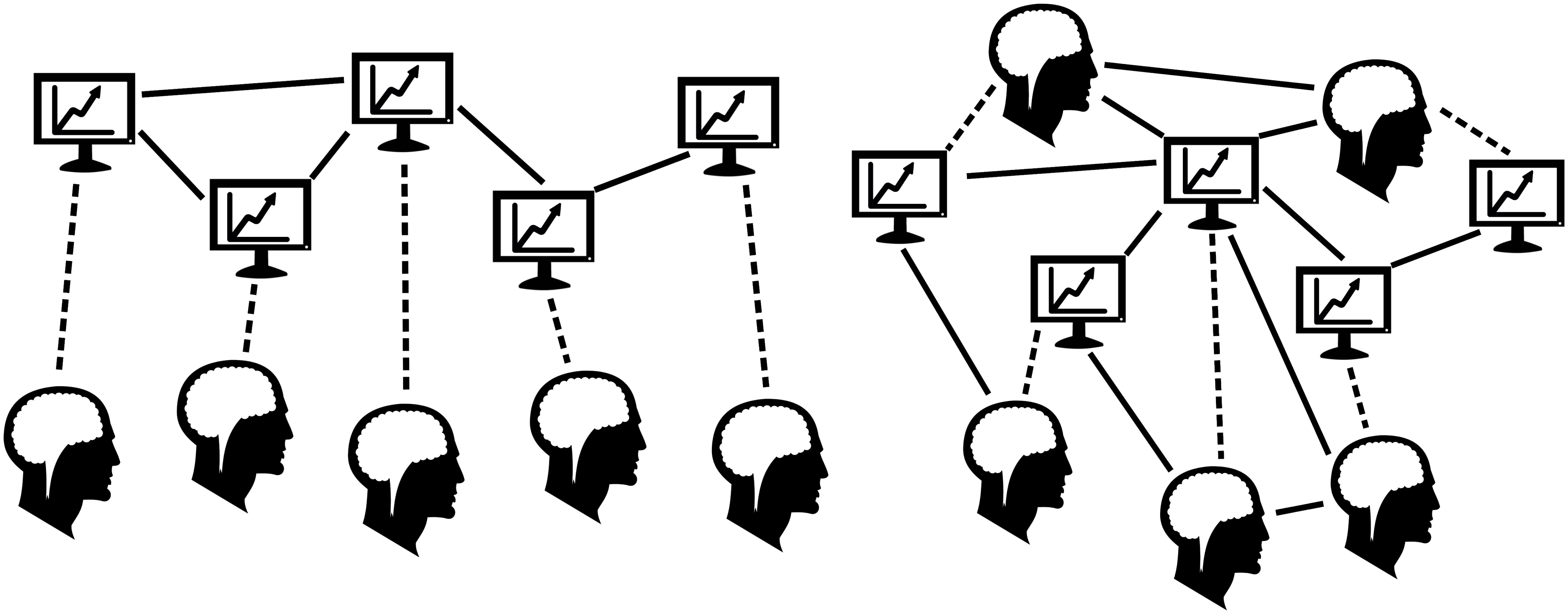
Traffic 'Shockwaves'

'Flash Crashes'

‘Intelligent’ Socio-Technical Systems?



'Intelligent' Socio-Technical Systems?



Computational Self-Awareness

Humans are able to figure many of these things out:

- We reason through situations, updating **our understanding of the world and our role in it**, as we go.
- We **reflect on our own behaviour from a social perspective**, and modify it accordingly.

$\longrightarrow$

Key Idea:

Systems that engage in **self-awareness** can:

- **better manage trade-offs** between goals at run-time, in complex, uncertain and dynamic environments, and
- **better fit in to society**, such that they are sensitive to laws, norms, ethics, expectations, and their own impact.

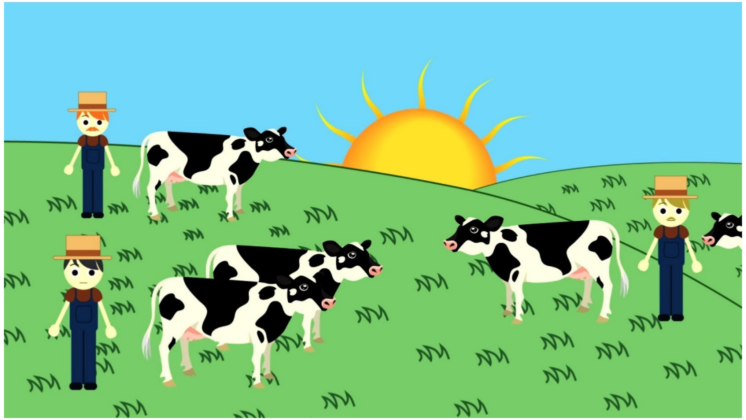
[5] Lewis, PR, Chandra A, Faniyi F, Glette K, Chen T, Bahsoon R, et al. **Architectural Aspects of Self-aware and Self-expressive Computing Systems**. IEEE Computer. 2015;48:62–70

[6] Lewis, PR, Platzner M, Rinner B, Torresen J, Yao X, editors. **Self-Aware Computing Systems: An Engineering Approach**. Springer; 2016

[7] Lewis, PR. **Self-aware Computing Systems: From Psychology to Engineering**. In: Proceedings of the 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE); 2017. p. 1044–1049

An Example: Collective Action Problems

- Individuals 'should' cooperate by contributing their time, effort and resources to a **shared enterprise**.
- Contributing comes at a cost: battery life, data usage, personal data, time, ...
- These systems risk being subject to the Tragedy of the Commons.
- Rational Choice Theory predicts that individuals act to maximise their own payoff at the **expense of the group**.
- **Why would people contribute to these systems?**



Avoiding the Tragedy of the Commons

- Ostrom's [8] fieldwork, observed that in many cases, societies are able to avoid anti-social outcomes by devising rules – **social expectations** – that govern their interactions. Examples: fisheries, irrigation systems, remote villages.
- **'Institutions'.**
- The Tragedy assumes that we are playing a particular game form, n-player single shot Prisoner's Dilemma.

[8] Ostrom E. **Governing the Commons : The Evolution of Institutions for Collective Action**. Cambridge University Press; 1990

- Humans using their cognitive and social abilities to reason through the game they find themselves in, and **collectively choosing to change the 'rules of the game' to something more advantageous**.
- This requires sustained cooperation,
  through **awareness of and compliance with the established social expectations**.

Avoiding the Tragedy of the Commons

- Ostrom's [8] fieldwork, observed that in many cases, societies are able to avoid anti-social outcomes by devising rules – **social expectations** – that govern their interactions. Examples: fisheries, irrigation systems, remote villages.
- **'Institutions'.**
- The Tragedy assumes that we are playing a particular game form, n-player single shot Prisoner's Dilemma.

[8] Ostrom E. **Governing the Commons : The Evolution of Institutions for Collective Action**. Cambridge University Press; 1990

- Humans using their cognitive and social abilities to reason through the game they find themselves in, and **collectively choosing to change the 'rules of the game' to something more advantageous**.
- This requires sustained cooperation,
  through **awareness of and compliance with the established social expectations**.
- **So why would autonomous systems decide to contribute?**
- **What sort of self-awareness would they need to be able to decide in a similar way to humans?**

Social Self-Awareness for Socio-Technical Systems

Socio-technical systems $+$ self-awareness $\longrightarrow$ social self-awareness.

Social Self-Awareness for Socio-Technical Systems

Socio-technical systems + self-awareness $\longrightarrow$ social self-awareness.

Social self-awareness:

- Ability to judge ones own behavior from a social perspective, to identify a social standard and compare oneself against it [9].

- Within animals too, we see individuals within a group that understand:
  - their position in the group
  - their relation to particular others
  - what behavioural expectations follow from these factors [10].

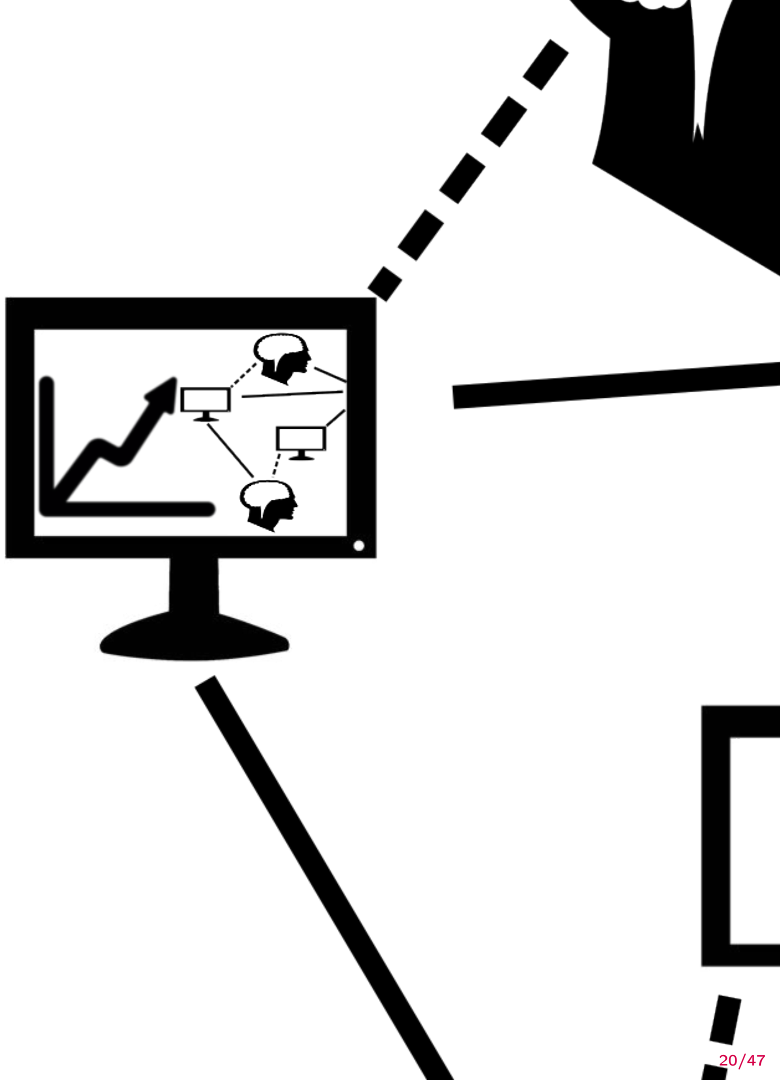
[9] Diener E, Srull TK. **Self-awareness, psychological perspective, and self-reinforcement in relation to personal and social standards**. Journal of Personality and Social Psychology. 1979;37:413–423

[10] DeGrazia D. In: Lurz R, editor. **Self-Awareness in Animals**. Cambridge University Press; 2009. p. 201–217

Towards Socially Self-Aware Machines

"An internal model allows a system to look ahead to the future consequences of current actions, without actually committing itself to those actions." [11, p25]

[11] Holland JH. **Complex Adaptive Systems**. Daedalus; 1992

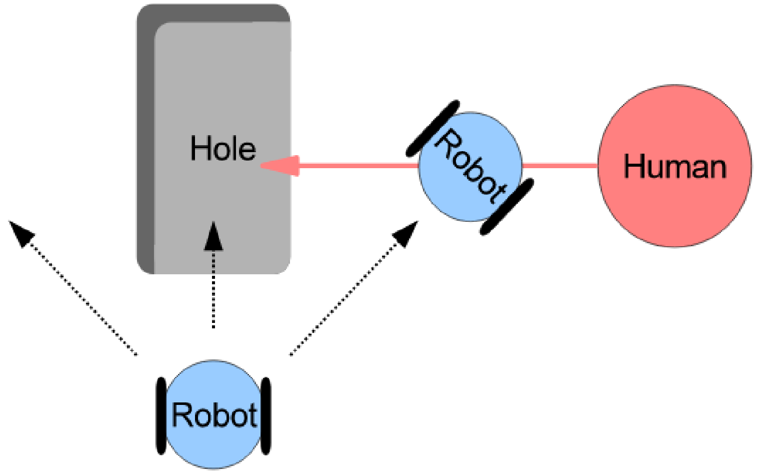


Key Research Questions:

- Can we design autonomous systems that use social models, at run-time, to reflect on:
  - their congruence with social expectations; and
  - the long-term social impact of their actions?
- What sort of models could we use?

Models for Social Self-Awareness

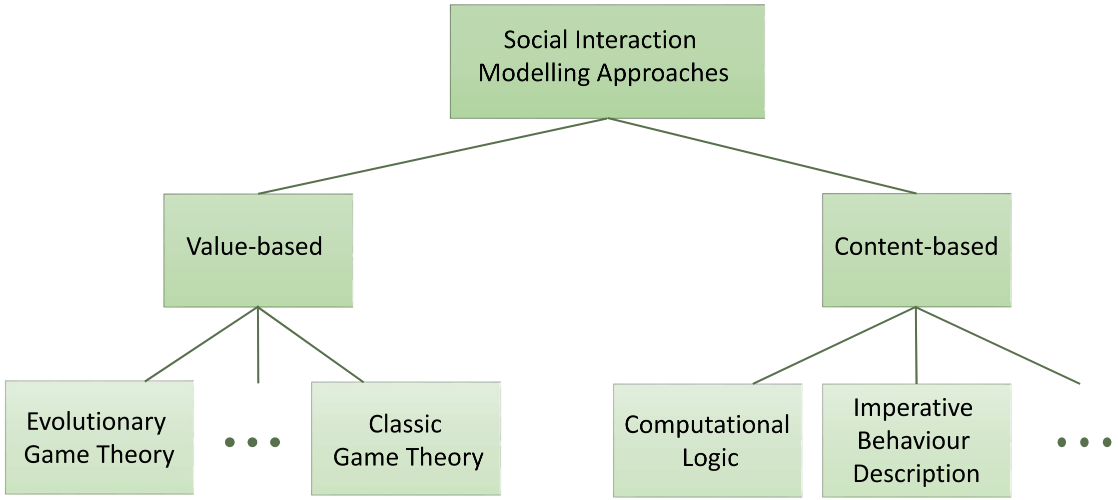
Example: Winfield's Safe and Ethical Robots [12].



- Models as *'consequence engines'* for autonomous systems.
- Simulation of self and others, to inform decision making.
- Dennett's tower of *Generate-And-Test*: how is survivability (fitness) impacted by the sophistication of mechanisms for generating and testing hypotheses about how to react in the world?
- Safe and ethical autonomous robots may not be achievable at all without mechanisms for self-awareness.

[12] Winfield A. **Robots with Internal Models: A Route to Self-Aware and Hence Safer Robots**. In: Pitt J, editor. The Computer After Me: Awareness and Self-awareness in Autonomic Systems; 2014. p. 237–252

Content-Based vs Value-Based Modelling



Two families of approaches to modelling social systems, which:

- Come from different perspectives in what they capture,
- Have different explanatory and analytical power.

[13] Powers ST, Ekárt A, Lewis, PR. **Modelling Enduring Institutions: The Complementarity of Evolutionary and Agent-based Approaches**. Cognitive Systems Research. 2018;52:67–81

Sustaining Cooperative Collective Action through Institutions

Question:

When will group members put effort into maintaining an institution, at a personal cost to themselves, so that it continues to provide beneficial outcomes for the group?

Sustaining Cooperative Collective Action through Institutions

Question:

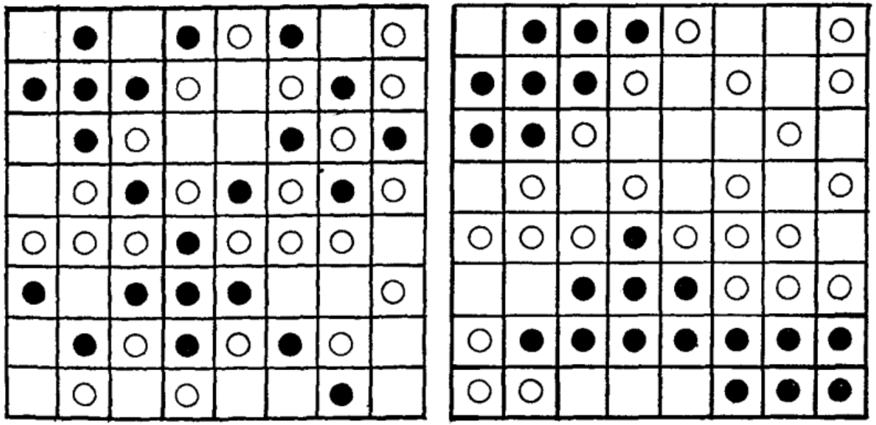
When will group members put effort into maintaining an institution, at a personal cost to themselves, so that it continues to provide beneficial outcomes for the group?

- Creating rules, **monitoring** them, and **sanctioning** rule-breakers requires time and effort (agent take on roles).
- Without this they collapse: individuals revert back to the default game form where cooperation is not favoured.
- To predict if institutions endure, we must examine **incentives for agents to take on institutional roles**.

- Previous studies used agent-based models, based on executing the **content** of individuals' behaviours.
- Here, I will show how this can be complemented by evolutionary game theory, based on the **value** of strategies.
- This enables us to provide new insight into:
  - The effects of different ways of incentivising individuals to take on an institutional role,
  - The optimal proportion of a group's resources that they should invest into supporting the institution, and
  - How to 'bootstrap' an institution / cooperation.

Agent-Based Modelling

- Agent-based modelling is a common technique for exploring social and economic systems.
- *Digital Petri Dishes* [14].
- Enable us to explore the consequences of a set of **premises**, under some **assumptions**.
- Method: explore empirically, claims via induction.

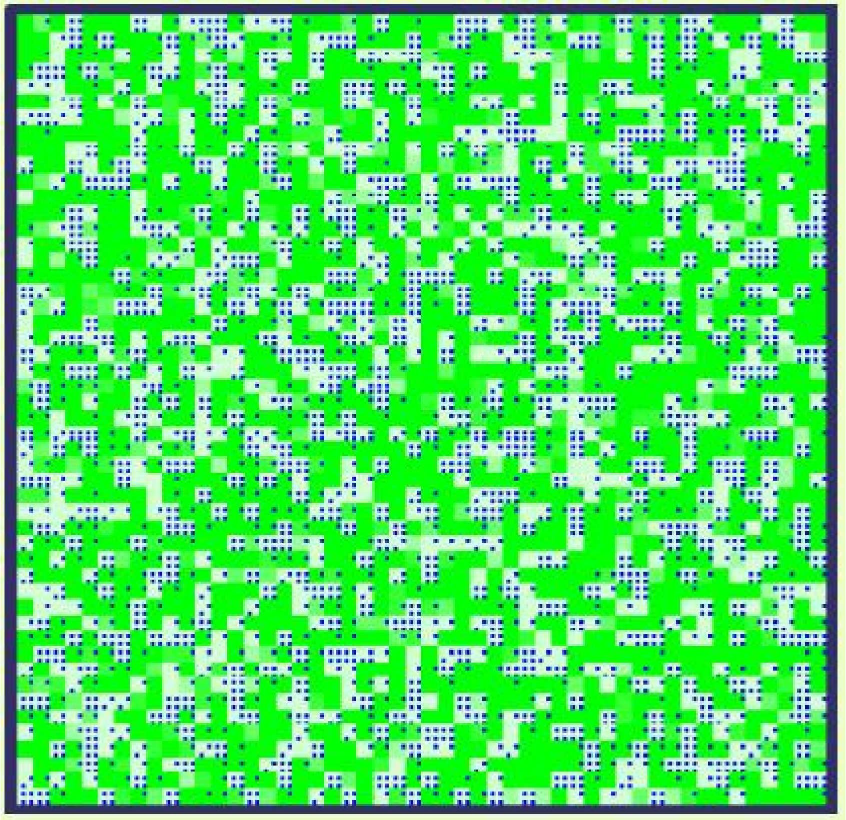


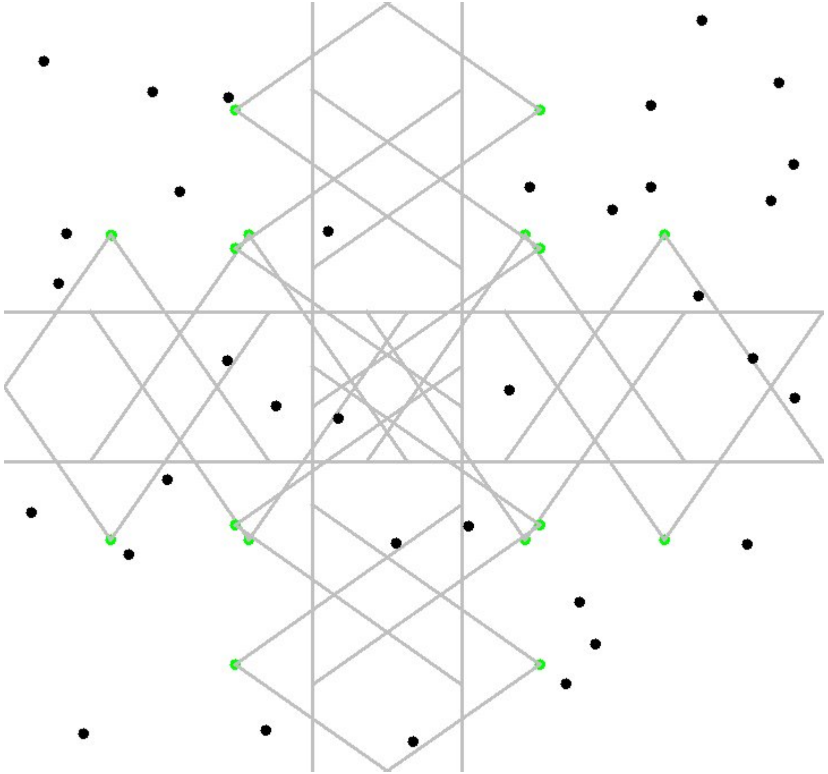
(a) (b)

Source: SCHELLING, T .C. (1974) On the ecology of micromotives. The Corporate society. Marris, R. (ed). London: Macmillan: 19-64.

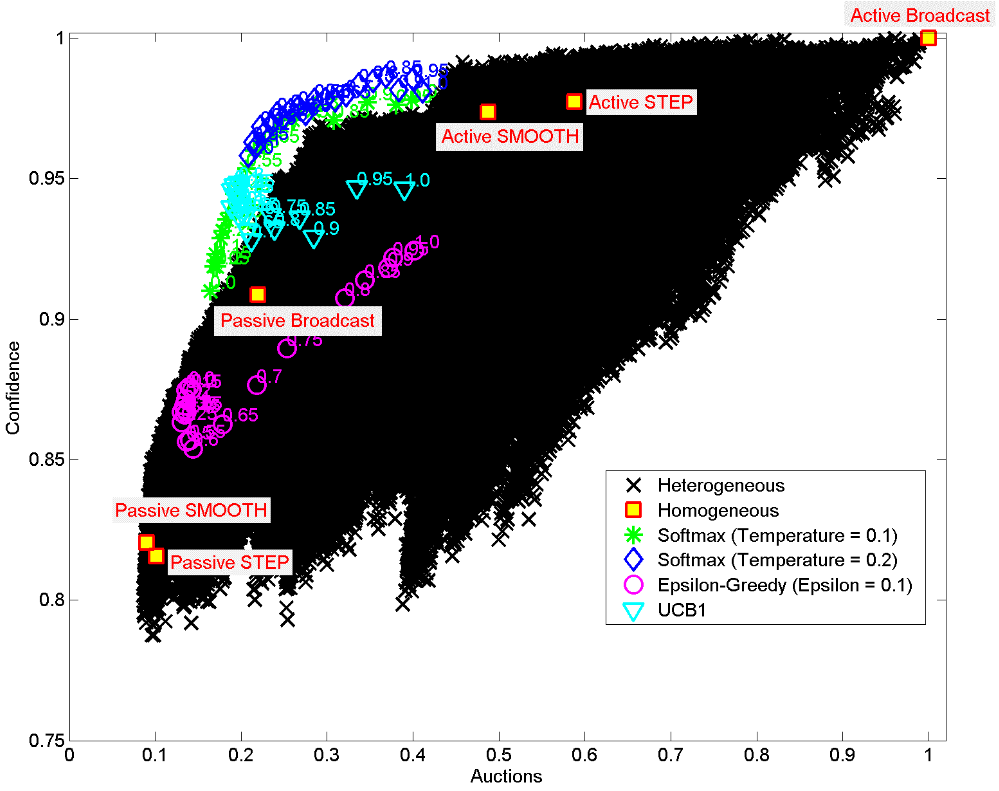
[14] Gavin M. **Agent-Based Modeling and Historical Simulation**. Digital Humanities Quarterly. 2014;8(4)

Agent-Based Modelling





Agent-Based Modelling Is Time Consuming!



[15] Lewis, PR, Esterle L, Chandra A, Rinner B, Torresen J, Yao X. **Static, Dynamic, and Adaptive Heterogeneity in Distributed Smart Camera Networks**. ACM Transactions on Autonomous and Adaptive Systems. 2015 Jun;10(2):8:1–8:30

Agent-Based Modelling of Institutions

Ostrom's principles for sustainable institutions have been formalised and operationalised by Pitt [16], in order to ask questions about when institutions endure.

Some key **results on institutions obtained by agent-based modelling**:

- The first three of Ostrom's principles are sufficient when agents comply with appropriation rules, otherwise the first six are necessary [16].
- The way the agents learn influences directly the existence and sustainability of the institution, and at the same time, the institution's features can either tolerate or inhibit learning [17].
- Relaxation and principled violation of norms promotes sustainability of institutions [18, 19].

[16] Pitt J, Schaumeier J, Artikis A. **Axiomatization of Socio-economic Principles for Self-organizing Institutions: Concepts, Experiments and Challenges**. ACM Transactions On Autonomous And Adaptive Systems. 2012;7

[17] Lewis, PR, Ekárt A. **Social and Asocial Learning in Collective Action Problems: The Rise and Fall of Socially-Beneficial Behaviour**. In: 2nd IEEE International Workshops on Foundations and Applications of Self* Systems, FAS*W 2017; 2017. p. 91–96

[18] Kurka DB, Pitt J. **The Principled Violation of Policy: Norm Flexibilization in Open Self-Organising Systems**. In: 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W); 2017. p. 33–38

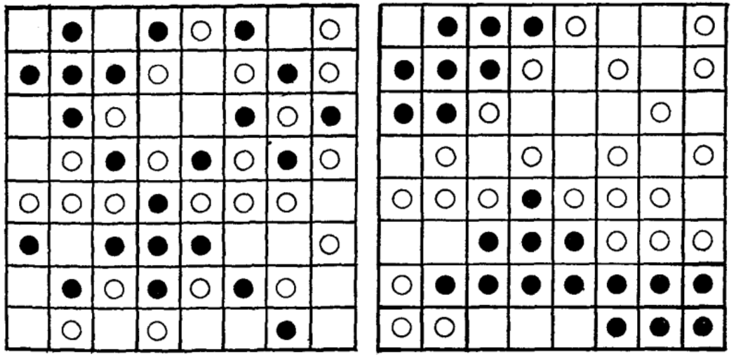
[19] Kurka DB, Pitt J, Lewis, PR, Patelli A, Ekárt A. **Disobedience as a Mechanism of Change**. In: Proceedings of the 12th IEEE International Conference on Self-adaptive and Self-organising Systems; 2018. p. 1–10 – ***Best Paper Award Winner.***

Attraction of Agent-Based Modelling

- Very common and attractive approach to take.
- The ability to **capture complex behaviours** and interactions in executable form, and to explore emergent phenomena simply by ‘running’ variants of the model.

Strengths:

- Particularly helpful when building intuition, establishing possibility results and plausibility, or illustrating counter-examples.
- Naturally capture spatial and temporal factors.
- Easy to build and prototype with, from a behavioural description.



(a) (b)

But:

- Complexity of the formal description limits its explanatory power.
- This is particularly true when answering questions related to incentivisation and critical values of parameters in a precise way.

Evolutionary Game Theoretic Modelling

Evolutionary game theory [20]:

- A more descriptive modelling technique, that **captures the evolutionary value of a strategy**.
- First established in theoretical biology to study the evolution of adaptive traits in populations of animals.
- Since been applied in economics, sociology, anthropology, and elsewhere in biology.
- Used to explore both genetic and cultural evolution.

Assumptions:

- We can interpret the utility as evolutionary fitness.
- Strategies replicate proportionally to their fitness.

Using Evolutionary Game Theory:

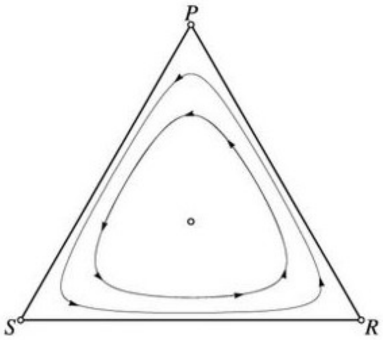
- Capture the **value** of a strategy, rather than its **content**.
- Modeller's task: to write equations that describe the fitness (i.e. evolutionary value) of each strategy, given the current frequency of each possible strategies in the population.

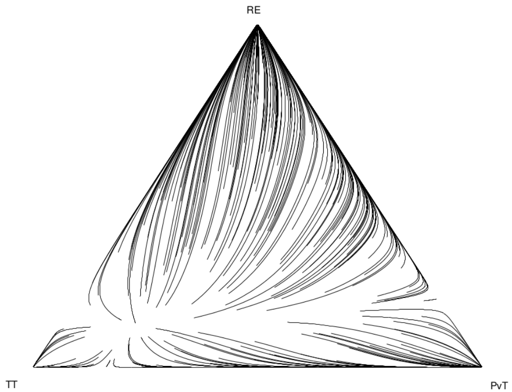
[20] Maynard Smith J. **Evolution and the Theory of Games**. Cambridge University Press; 1982

Performing Analysis with Evolutionary Game Theory

Typically evolutionary game theory looks for:

- **Evolutionary stable strategies**, a game state where a mutant strategy cannot successfully enter the population to disturb the existing dynamic. ESSs are therefore stable for the long term (until the game changes).
- **Population dynamics**, i.e. predictable patterns of trait frequency change over time, critical thresholds.





No requirement to 'run' the model, equations are either:

- solved analytically
- solved numerically by computer, when tractability is an issue,
- explored through discrete 'individual-based simulations'

Aim is to **deduce relationships and critical parameter values**.

Building an Evolutionary Game Theoretic Model of Institutions

n agents take part in a linear public goods game. Each agent makes two decisions:

1. $\iota_{i\mathrm{C}}$: **whether or not to cooperate** by contributing to the common pool at a cost to itself (total number $= n_{\mathrm{C}}$);
2. $\iota_{i\mathrm{M}}$: **whether or not to monitor** other agents to determine if they have contributed.

B_{G} represents the individual's share of the common-pool resource, computed as:

$$B_{\mathrm{G}} = \frac{1}{n} \times \alpha\, n_{\mathrm{C}}, \tag{1}$$

Costs and Rewards:

- Contributing to the common-pool resource comes at cost $C_C < \alpha$. Hence $\alpha n_C > C_C n_C$.
- Agents that did not contribute and were monitored (thus caught) are sanctioned, creating a cost to free-riding, C_F.
- Monitoring carries some cost to the agent, C_M.
- Monitors are reimbursed for their work, B_M, according to a reimbursement scheme (more later).

Therefore, the utility of agent i is given by

$$u_i = B_{\mathrm{G}} - \iota_{i\mathrm{C}} C_{\mathrm{C}} - (1 - \iota_{i\mathrm{C}}) C_{\mathrm{F}} + \iota_{i\mathrm{M}} [B_{\mathrm{M}} - C_{\mathrm{M}}]. \tag{2}$$

Based on Powers ST, Ekárt A, Lewis, PR. **Modelling Enduring Institutions: The Complementarity of Evolutionary and Agent-based Approaches**. Cognitive Systems Research. 2018;52:67–81

Costs of Freeriding and Monitoring

C_{F} **represents the cost of free-riding**, paid by all agents with $\iota_{\mathrm{C}} = 0$:

$$C_{\mathrm{F}} = \frac{p\,n_{\mathrm{M}}}{n} s, \tag{3}$$

where:

- n_{M} is the number of agents that take on the monitoring role, i.e. the number of agents with $\iota_{\mathrm{M}} = 1$, and
- p is the number of agents monitored by each monitor.

I.e. the probability than an agent is monitored, multiplied by the **sanction imposed if detected free-riding**, s.

C_{M} **represents the cost to the agent of monitoring** other agents.

The cost of monitoring a single agent is δ, so the total cost to an agent of monitoring on one round is

$$C_{\mathrm{M}} = p\delta. \tag{4}$$

Incentivising Cooperation

We are interested in the conditions under which **agents will create a system of monitoring that incentivises cooperation**.

I.e. when is $C_{\mathrm{C}} < C_{\mathrm{F}}$?

To determine this, we consider the evolution of the two agent behavioural traits:

- ι_{C}, whether you contribute
- ι_{M}, whether you monitor

when agents with those traits are in competition with each other.

When Will Cooperation be an Equilibrium?

Monitors are paid from the common-pool resource according to an **‘investment proportion’** parameter β.

β is an **institutional fact** that agents must collectively decide through a political game.

When Will Cooperation be an Equilibrium?

Monitors are paid from the common-pool resource according to an **'investment proportion'** parameter β.

β is an **institutional fact** that agents must collectively decide through a political game.

Thus, we now have

$$B_{\mathrm{G}} = (1-\beta)\frac{1}{n} \times \alpha\, n_{\mathrm{C}} \tag{5}$$

When Will Cooperation be an Equilibrium?

Monitors are paid from the common-pool resource according to an **'investment proportion'** parameter β.

β is an **institutional fact** that agents must collectively decide through a political game.

Thus, we now have

$$B_{\mathrm{G}} = (1-\beta)\frac{1}{n} \times \alpha\, n_{\mathrm{C}} \tag{5}$$

Analytically, we can determine that for full cooperation (i.e. setting $\frac{n_{\mathrm{C}}}{n} = 1$), it must be that:

$$\frac{\alpha\beta}{\delta} > \frac{C_{\mathrm{c}}}{s}. \tag{6}$$

When we choose our institutional facts $(s,\ \beta)$ such that this holds, **full cooperation will be an equilibrium**.

How Much Resource Should be Devoted to Monitoring?

We can now also **determine the minimum value of β** (how much resource to devote to paying for monitoring) necessary to make full cooperation an equilibrium.

By rearrangement, we obtain:

$$\beta > \frac{\delta C_{\mathrm{C}}}{\alpha s}. \tag{7}$$

When this inequality holds, and the agents are all cooperating, then a sufficient level of monitoring is incentivised to maintain full cooperation.

How Much Resource Should be Devoted to Monitoring?

We can now also **determine the minimum value of** β (how much resource to devote to paying for monitoring) necessary to make full cooperation an equilibrium.

By rearrangement, we obtain:

$$\beta > \frac{\delta C_{\mathrm{C}}}{\alpha s}. \tag{7}$$

When this inequality holds, and the agents are all cooperating, then a sufficient level of monitoring is incentivised to maintain full cooperation.

Thus,
Q: how much of their resources should a group invest into monitoring?
A: The smallest value of β **that satisfies this** (more is wasteful).

This highlights how value-based models can produce **precise predictions** about how to control a system.

When Will Agents Reach this Equilibrium from Non-Cooperation?

Q: How much monitoring is necessary to incentivise cooperation when there are no cooperators in the group?

- It can be shown that monitoring is not incentivised when $\frac{n_{\mathrm{C}}}{n} = 0$, i.e. when no agents are currently cooperating.
- Consequently, **no cooperation and no monitoring is also an equilibrium.**
- Yet, this represents a natural starting point when considering the origin of institutions.

When Will Agents Reach this Equilibrium from Non-Cooperation?

Q: How much monitoring is necessary to incentivise cooperation when there are no cooperators in the group?

- It can be shown that monitoring is not incentivised when $\frac{n_{\mathrm{C}}}{n} = 0$, i.e. when no agents are currently cooperating.
- Consequently, **no cooperation and no monitoring is also an equilibrium.**
- Yet, this represents a natural starting point when considering the origin of institutions.

Q: How might a group break free from this *anti-social* equilibrium and move to the *cooperate-and-monitor* equilibrium where everyone is better off?

When Will Agents Reach this Equilibrium from Non-Cooperation?

Q: How much monitoring is necessary to incentivise cooperation when there are no cooperators in the group?

- It can be shown that monitoring is not incentivised when $\frac{n_C}{n} = 0$, i.e. when no agents are currently cooperating.
- Consequently, **no cooperation and no monitoring is also an equilibrium.**
- Yet, this represents a natural starting point when considering the origin of institutions.

Q: How might a group break free from this *anti-social* equilibrium and move to the *cooperate-and-monitor* equilibrium where everyone is better off?

- It requires some agents to initially monitor for free (i.e., carry out pro-social behaviour), after which the cooperative, higher-payoff equilibrium will be reached.

When Will Agents Reach this Equilibrium from Non-Cooperation?

Q: How much monitoring is necessary to incentivise cooperation when there are no cooperators in the group?

- It can be shown that monitoring is not incentivised when $\frac{n_{\mathrm{C}}}{n} = 0$, i.e. when no agents are currently cooperating.
- Consequently, **no cooperation and no monitoring is also an equilibrium.**
- Yet, this represents a natural starting point when considering the origin of institutions.

Q: How might a group break free from this *anti-social* equilibrium and move to the *cooperate-and-monitor* equilibrium where everyone is better off?

- It requires some agents to initially monitor for free (i.e., carry out pro-social behaviour), after which the cooperative, higher-payoff equilibrium will be reached.

However,

- Because we are describing a behaviour that requires agents to be forward-looking, **it cannot be fully captured in an evolutionary game theory model** where individuals' cognition is completely myopic.
- But it could be explored in a content-based model that implements cognitive theories of agent behaviour.

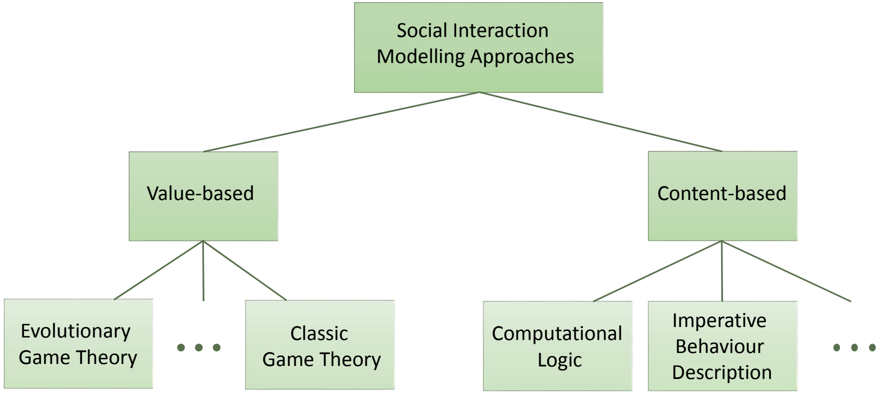
So What Have We Learnt?

There is an important distinction between content-based and value-based modelling approaches.

Each leaves implicit what is made explicit in the other:

- **ABM captures a large and complex set of behaviours** (e.g. learning, deliberation, meta-cognition), and the effect they have on the world and other agents, but struggles to support an analysis of their value.

- **EGT provides the primitives to analyse incentives associated with behaviours**, yet the (complex) nature of the strategies that lead to these values is omitted. It is assumed that value is accurately defined.



Questions That Remain to Realise Social Self-Awareness

How suitable are these approaches for run-time modelling, to achieve social self-awareness?

- Can they capture the required cognition and behaviour (**sufficient theory of mind**)?
- Do they provide sufficient value (**sufficient predictive power**)?
- How can self-aware systems **induce these models from experience**?
- What mechanisms can self-aware systems use to **reason about these models** in an automated way?
- How do they **scale**, with respect to resource constraints and decision time?

Questions That Remain to Realise Social Self-Awareness

How suitable are these approaches for run-time modelling, to achieve social self-awareness?

- Can they capture the required cognition and behaviour (**sufficient theory of mind**)?
- Do they provide sufficient value (**sufficient predictive power**)?
- How can self-aware systems **induce these models from experience**?
- What mechanisms can self-aware systems use to **reason about these models** in an automated way?
- How do they **scale**, with respect to resource constraints and decision time?

Hybridising ABM and EGT?

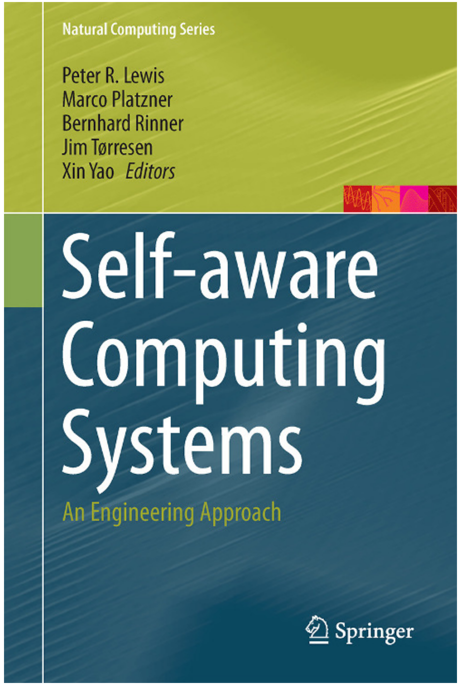
- Can we **induce the value of behaviours empirically**, from a content-based model? Use machine learning?
- Is it easier to **extract value** from a content-based model or to **add behavioural content** to a value-based one?
- Are there behaviours that defy value-based modelling?

Where Next?

P. R. Lewis, M. Platzner, B. Rinner, J. Torresen, and X. Yao, editors.
Self-Aware Computing Systems: An Engineering Approach.
Springer, 2016

P. R. Lewis.
Self-aware Computing Systems: From Psychology to Engineering.
In Proceedings of the 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE), pages 1044–1049, 2017

S. T. Powers, A. Ekárt, and P. R. Lewis.
Modelling Enduring Institutions: The Complementarity of Evolutionary and Agent-based Approaches.
Cognitive Systems Research. 2018; 52:67–81



p.lewis@aston.ac.uk

References I

[1] Chitty DM, Wanner E, Parmar R, Lewis, PR.
Scaling ACO to large-scale vehicle fleet optimisation via Partial-ACO.
In: GECCO '19: Proceedings of the Genetic and Evolutionary Computation Conference Companion. ACM; 2019. .

[2] Chitty DM, Parmar R, Lewis, PR.
Improving Urban Air Quality Through Long-Term Optimisation of Vehicle Fleets.
Advances in Intelligent Systems and Computing. Springer; 2019. .

[3] Bellman KL, Botev J, Hildmann H, Lewis, PR, Marsh S, Pitt J, et al.
Socially-Sensitive Systems Design: Exploring Social Potential.
IEEE Technology and Society Magazine. 2017;36(3):72–80.

[4] Andras P, Esterle L, Guckert M, Han TA, Peter R Lewis, Milanovic K, et al.
Trusting Intelligent Machines.
IEEE Technology and Society Magazine. 2018;37.

[5] Lewis, PR, Chandra A, Faniyi F, Glette K, Chen T, Bahsoon R, et al.
Architectural Aspects of Self-aware and Self-expressive Computing Systems.
IEEE Computer. 2015;48:62–70.

References II

[6] Lewis, PR, Platzner M, Rinner B, Torresen J, Yao X, editors.
Self-Aware Computing Systems: An Engineering Approach.
Springer; 2016.

[7] Lewis, PR.
Self-aware Computing Systems: From Psychology to Engineering.
In: Proceedings of the 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE); 2017. p. 1044–1049.

[8] Ostrom E.
Governing the Commons : The Evolution of Institutions for Collective Action.
Cambridge University Press; 1990.

[9] Diener E, Srull TK.
Self-awareness, psychological perspective, and self-reinforcement in relation to personal and social standards.
Journal of Personality and Social Psychology. 1979;37:413–423.

[10] DeGrazia D.
In: Lurz R, editor. **Self-Awareness in Animals**. Cambridge University Press; 2009. p. 201–217.

References III

[11] Holland JH.
Complex Adaptive Systems.
Daedalus; 1992.

[12] Winfield A.
Robots with Internal Models: A Route to Self-Aware and Hence Safer Robots.
In: Pitt J, editor. The Computer After Me: Awareness and Self-awareness in Autonomic Systems; 2014. p. 237–252.

[13] Powers ST, Ekárt A, Lewis, PR.
Modelling Enduring Institutions: The Complementarity of Evolutionary and Agent-based Approaches.
Cognitive Systems Research. 2018;52:67–81.

[14] Gavin M.
Agent-Based Modeling and Historical Simulation.
Digital Humanities Quarterly. 2014;8(4).

[15] Lewis, PR, Esterle L, Chandra A, Rinner B, Torresen J, Yao X.
Static, Dynamic, and Adaptive Heterogeneity in Distributed Smart Camera Networks.
ACM Transactions on Autonomous and Adaptive Systems. 2015 Jun;10(2):8:1–8:30.

References IV

[16] Pitt J, Schaumeier J, Artikis A.
Axiomatization of Socio-economic Principles for Self-organizing Institutions: Concepts, Experiments and Challenges.
ACM Transactions On Autonomous And Adaptive Systems. 2012;7.

[17] Lewis, PR, Ekárt A.
Social and Asocial Learning in Collective Action Problems: The Rise and Fall of Socially-Beneficial Behaviour.
In: 2nd IEEE International Workshops on Foundations and Applications of Self* Systems, FAS*W 2017; 2017. p. 91–96.

[18] Kurka DB, Pitt J.
The Principled Violation of Policy: Norm Flexibilization in Open Self-Organising Systems.
In: 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W); 2017. p. 33–38.

[19] Kurka DB, Pitt J, Lewis, PR, Patelli A, Ekárt A.
Disobedience as a Mechanism of Change.
In: Proceedings of the 12th IEEE International Conference on Self-adaptive and Self-organising Systems; 2018. p. 1–10.

[20] Maynard Smith J.
Evolution and the Theory of Games.
Cambridge University Press; 1982.

Ostrom's Principles

Ostrom proposed eight principles that, when present, led to institutions enduring.

1. Clearly defined boundaries
2. Congruence between appropriation and provision rules and local conditions
3. Collective-choice arrangements: Appropriators can participate in the design of the institution by tailoring the rules over time.
4. Monitoring
5. Graduated sanctions
6. Conflict-resolution mechanisms.
7. Minimal recognition of rights to organize.
8. Nested enterprises.

Ostrom predicts that where these eight principles are satisfied institutions will be maintained and will continue to prevent over-exploitation of common- pool resources over a long time horizon.

We focus on monitoring in this work.

Some General Recommendations

Aimed at those using modelling approaches to understand, control and design social and socio-technical systems.

Recommendation 1: Use both content-based and value-based approaches

- Previous results obtained from agent-baed techniques, can be complemented with those obtained by taking an evolutionary game theoretic approach.
- The results presented in this talk would have been difficult to obtain empirically.
- However, the evolutionary approach would struggle to obtain results concerning the interactions of more complex cognitive agent behaviours, such as those associated with forward looking behaviour and richer human social interactions (cf. Ostrom).

Some General Recommendations

Aimed at those using modelling approaches to understand, control and design social and socio-technical systems.

Recommendation 1: Use both content-based and value-based approaches

- Previous results obtained from agent-baed techniques, can be complemented with those obtained by taking an evolutionary game theoretic approach.
- The results presented in this talk would have been difficult to obtain empirically.
- However, the evolutionary approach would struggle to obtain results concerning the interactions of more complex cognitive agent behaviours, such as those associated with forward looking behaviour and richer human social interactions (cf. Ostrom).

Recommendation 2: Don't worry if the value-based model is not complete

- It is tempting to think that without being complete, a value-based model has limited value.
- Yes, more complete models lead to more complete results, but even partial models can expose insightful inequalities.
- E.g. we have not considered selection on cooperation and monitoring at the same time, which would lead to statements on their co-variance.

Some General Recommendations

Recommendation 3: Go for the qualitatively equivalent, but more tractable alternative

- Often seemingly innocuous changes to a model can drastically change the tractability of value-based models.
- E.g. we assumed that each monitor perfectly coordinates to monitor a non-overlapping set of agents. If overlaps are allowed, the proportion of agents monitored no longer increases linearly with the number of monitors.

  $\frac{n_{\mathrm{M}}}{n} > \frac{C_{\mathrm{c}}}{ps}$ becomes $C_{\mathrm{F}}(t) = [1 - (1 - \frac{p}{n})^{n_{\mathrm{M}}(t)}]s$

  which is substantially harder to solve to determine the required level of monitoring.

Some General Recommendations

Recommendation 3: Go for the qualitatively equivalent, but more tractable alternative

- Often seemingly innocuous changes to a model can drastically change the tractability of value-based models.
- E.g. we assumed that each monitor perfectly coordinates to monitor a non-overlapping set of agents. If overlaps are allowed, the proportion of agents monitored no longer increases linearly with the number of monitors.

  $\frac{n_{\mathrm{M}}}{n} > \frac{C_{\mathrm{c}}}{ps}$ becomes $C_{\mathrm{F}}(t) = [1 - (1 - \frac{p}{n})^{n_{\mathrm{M}}(t)}]s$

  which is substantially harder to solve to determine the required level of monitoring.

Recommendation 4: Clarify your assumptions, not just premises

- Premises are what we are exploring the impact of, assumptions are introduced in how we conduct the study.
- It is important for modellers to provide clarity on whether:
  - models assume or explore the extent to which agents engage in cognition, or
  - if they assume that agents simply 'behave'.
- This context is important because predictions may vary drastically as a result.