Unleashing the Agents: From a Descriptive to an Explanatory Perspective in Agent-based Modelling

Christopher K. Frantz

Norwegian University of Science and Technology (NTNU), Norway
cf@christopherfrantz.org, christopher.frantz@ntnu.no

Abstract. Agent-based modelling endows the experimenter with high levels of flexibility, and consequently, responsibility. Possibly because of that, developing good models is hard. In this work, we engage in the discussion around improving the analytical value and disciplinary acceptance of agent-based social simulation. To this end, this paper includes the proposal to make the agents themselves observers, as opposed to just participants, of the simulation to introduce explanatory power that cannot be leveraged by on descriptive macro-level analysis alone. This is followed by an argument for the use of institutional concepts for any mechanism that seeks to embed quasi-reflective capabilities in an effort to gain accessible explanatory insights from simulations. To exemplify this idea we apply it to a cooperation game of moderate complexity, and finally discuss application opportunities, challenges and future directions.

Keywords: Agent-based Modelling, Social Simulation, Methodology, Explanatory Simulation, Institutions, Norms, Corruption Game, Institutional Analysis, Grammar of Institutions, Nested ADICO, Stereotyping, Implicit Social Cognition, Social Learning

1 Introduction

Agent-based Modelling and Simulation (ABM) (Gilbert 2008) is experiencing uptake for an increasingly wide range of coordination and cooperation problems based on its accessible agent metaphor and the ability to reconstruct problems incrementally and from the perspective of the problem domain. A specific opportunity arising from the application of ABM is its ability to inform modelling from a theory-driven perspective and/or based on existing empirical data (Tolk 2015), making ABM suitable both for abstract conceptual work as well as for concrete applications (e.g., simulating behaviour during emergencies (Pan et al. 2007)).

However, on the flip side, the flexibility of the agent concept can be problematic. On the one hand, the principles of agent-based modelling encourage experimenters to think in terms of the problem domain, and do not constrain them to selectively favour complexity of the problem domain or the embedded

agent concept. On the other hand, the flexibility of the agent concept allows for the encoding of agent behaviour on arbitrary levels of complexity, construed either as simple execution rules, or as complex architectures able to account for cognitive and social determinants of behaviour.

To manage the complexity of the resulting scenarios, ABM makes it convenient for experimenters to detach themselves from the underlying agent implementation, and rather *ascribe* agency (and potentially intentionality) to the modelled entities, and focus on the analysed (generally macro-level) phenomenon without taking agency on the micro level into account. Combined with the analytical focus on the problem domain, this leaves researchers at risk to perceive the underlying agent model (if not the entire simulation) as a black box and treat the produced results at face value during the ensuing interpretation.

The resulting inability to account for the introduced assumptions and abstractions, alongside other methodological concerns (which we discuss in Section 2), challenges the broader adoption of ABM in various disciplines and is frequently put forth in advocacy for methods that can rely on a comprehensive formalisation of the problem.

In this work, we explore how we can address this concern, and shift the agent itself back into the spotlight as a means to provide better explanatory insights into the model dynamics. To achieve this, we will, of course, need to introduce a necessary minimal set of assumptions about the agents, which is compatible with calls to endow agent-based models with stronger social-psychological capabilities (see e.g., Jager (2017)). To illustrate this aspect, we will explore this idea using a conceptual cooperation problem with moderate socio-structural complexity that emulates prototypical behaviours found in economic exchange, such as corruption.

The paper is structured as follows: Section 2 sets the scene that motivates the use of agent conceptions that exhibit explanatory functions (while driving the phenomenon of interest). In Section 3, we develop a candidate approach to leverage deeper insights into the dynamics of agent-based models. Section 4 sketches a cooperation scenario that employs the proposed architecture, which is subsequently evaluated in Section 5. Section 6 concludes the paper with a discussion of the insights and outlines further research directions.

2 Background

The principles of social simulation more generally, and agent-based modelling specifically, have come a long way. Since Schelling's experimentation with cellular automata to analyse sociological phenomena (Schelling 1971) – marking the birth of social simulation – , Axelrod's seminal work on cooperation (Axelrod 1986) shifted agent-based concepts into the mainstream, and Epstein's declaration of simulation as the 'third way of doing science' (Epstein 1999) marked the methodological rite of passage. The accessibility of the intuitions underlying the

agent concept, the availability of de facto standard modelling platforms[1], and increasing maturity of methodological prescriptions and documentation standards (e.g., ODD+D (Müller et al. 2013)), lowered the threshold for the use of agent-based simulation in a wide range of disciplines, including political science (Cederman 2005), economics (Farmer and Foley 2009), institutional analysis (Frantz et al. 2014), social psychology (Jackson et al. 2017), criminology (Birks et al. 2012), and religious violence (Shults et al. 2017), to name a few.

Beyond the use as a tool for the analysis of specific phenomena, the principles of agent-based modelling have contributed to the exploration of fundamental sociological concepts, such as the role of trust for cooperation, social functions of reciprocity, and the influence of topology on opinion formation and dynamics.[2]

However, confronted with the complexity of interaction on a social level, modellers are required to make strong assumptions about the underlying agent concept, as reflected in the KISS vs. KIDS discussion (Edmonds and Moss 2005). This includes the decision whether to model agents as primitive rule executors without any autonomy and prescribed social interaction, or to opt for richer agent architectures that account for cognitive and social capabilities of humans[3], the consideration of bounded rationality (Simon 1955) or the scenario-dependent situational adaptation of behavioural strategies (e.g., Janssen and Jager (1999)). At the same time, it is at the modeller's discretion to decide how interactions between individuals are represented (e.g., in comprehensive detail or as compound action), and to what extent agents can observe their physical and social environment (limited observation, noise[4]), as well as their ability to retain and access information.

This flexibility, the seemingly arbitrary choice of detail and subsumption of socio-cognitive functions by abstract architectures, made ABM subject to criticism, including the objection to high-level abstractions, choice of assumptions and their empirical support (Lengnick 2013), epistemological challenges in identifying causal relationships in the first place (Grüne-Yanoff 2009)[5], practitioners' concerns with the abstract representation of agency (Levy et al. 2016), and, last but not least, challenges from a methodological standpoint (Galán and Izquierdo 2005; Galán et al. 2017).

Despite these concerns, agent-based modelling provides conceptual riches and explorative potential for the analysis of social systems that few other techniques can offer. It builds on the human metaphor without carrying psychological burdens of actual humans (biases, unintended learning effects, questionnaire fatigue, etc.), the control of which makes empirical studies with human participants ex-

[1] See Kravari and Bassiliades (2015) for a comprehensive overview. Abar et al. (2017)'s survey provides a refined differentiation of platforms by application domains.

[2] Bianchi and Squazzoni (2015) collated an insightful overview that illustrates the impact of ABM on sociology.

[3] For an overview refer to Balke and Gilbert (2014).

[4] The importance of considering noise in the physical and social environment has been convincingly argued by Macy and Tsvetkova (2015).

[5] Equally noteworthy is the rebuttal of Grüne-Yanoff's argument by Elsenbroich (2012).

pensive and error-prone. Being freed from such limitations, we propose to move beyond making agents mere actors in the scenarios of interest, and exploit their psychologically impartial nature and deterministic properties to make the agents themselves quasi-reflective observers of the scenario. In doing so, we can endow agents with an *explanatory* role following the motto: "Don't tell me *what* you do, tell me *why* you do it."

However, before developing this proposal in greater detail in the following section, it is important to guard against potential misconceptions of the proposed approach.[6] The seasoned modeller may suggest that most agent-based modelling platforms, in fact, offer mechanisms that allow the runtime inspection of agent properties – an aspect that relates to the intuitions of this work.[7] However, while such functionality exists, it is a) generally intended to support the development process in order to debug agent properties (e.g., resource levels), and b) is focused on the situational state of the inspected entity. The approach put forth in the following sections qualitatively differs in that it provides richer statement representations that aim at reflecting the 'narrative' of the scenario from the perspective of an agent – targeting the experimenter, as opposed to the developer. The proposed approach emphasises a dynamic perspective that captures and condenses the interaction history in an intuitively accessible syntactic form over the conventional comparative-static approach applied in the step-wise inspection of agent state, the latter of which leaves it to the experimenter to *manually infer* associated agent behaviour.

3 Concept

The central challenge in augmenting agents with human-like reflective capabilities – while retaining *scenario independence* of the approach and affording a *lightweight and accessible interpretation* – is to identify a basis to deliberate about the cognitive assumptions for a quasi-reflective agent.

3.1 Institution as a cognitive basis

Informing this decision, we could allude to the superior human reasoning capabilities, and consequently favour concepts that emphasise deliberation abilities, such as represented by cognitive agent architectures. However, the focus on such risks misrepresenting the mechanisms that facilitate humans' functioning in social groups and would neglect subconscious processes dominating routine-based decision-making (see e.g., Kahneman (2013)). Instead, for a realistic baseline representation of social functioning, let us suggest that we primarily rely on fundamental mechanisms that make our social environment computable by allowing us to develop predictive capabilities that accommodate the bounds of rationality (Simon 1955), while being adaptive to changing social and situational circumstances – institutions. Institutions (North 1991; Hodgson 2006), stylised as

[6] At this stage, it is important to acknowledge the anonymous reviewers who provided valuable feedback for further refinement.

[7] Noteworthy examples include Swarm (Minar et al. 1996), MASON (Luke et al. 2005), NetLogo (Tisue and Wilensky 2004) and Repast (North et al. 2013).

the "rules of the game" (North 1990), are entrenched social behaviour, such as conventions (e.g., which side of the road to drive on), social norms (e.g., queueing for payment), and rules (e.g., traffic regulation, contracts), that are imposed by some authority or arise based on emergent behaviour (e.g., collective action) and are transmitted by socialisation. Essential characteristics for functioning institutions are their adoption, accepted normative status, and subsequent embedding in participants' mental structure. This fundamental role of institutions becomes clearer when interpreting their establishment itself as self-referential, in that the "essence of belief is the establishment of habit" (Peirce 1878). Searle (2005) likewise deems institutional structures fundamentally embedded in our cognitive processes, and, assuming a more radical position, Castelfranchi suggests we can interpret "minds [themselves] as social institutions" (Castelfranchi 2014).

In essence, if we assume that the belief in institutions (irrespective of the concrete form) is the lowest common denominator of any individual's (and, in extension, any society's) belief system, the use of institution representations is a sensible starting point for leveraging the explanatory power of agents.

While we briefly discussed the role of institutions as fundamental structure, we have yet to clarify the relevant processes that we assume for the associated agent model. One of those is the concept of "implicit social cognition" (Greenwald et al. 2002), visible in the ability to form and operate on observed patterns of individual and social characteristics – a specific function we commonly refer to as *stereotyping*. This implies the ability to draw generalisations across multiple attribute combinations, something we humans are specifically good at. More importantly, we are fast to do so (Zeithamova et al. 2012), and willingly sacrifice accuracy and ignore representativeness. Another relevant function to understand and generalise social information is the ability to not only learn directly from personal experiences (*experiential learning*), but to learn from one's social environment by applying some form of *social learning* (Bandura 1977). However, while we deem the ability to rely on stereotypes for heuristic purposes as essential for the processing of behavioural information, the ability to learn from the social environment introduces stronger assumptions about the agents' sensing abilities, and, in consequence, for simulation scenarios. It is for this reason that we consider this an optional component of such baseline architecture.

With this position in mind, we will turn to a candidate representation mechanism from the area of institutional modelling and analysis that allows us to integrate the fundamental processes described above.

3.2 Nested ADICO (nADICO) for endogenous inference of social institutions

When intending to provide a generic way to capture individuals' observations to infer its institutional function, we are, of course, subjected to a wide range of potential representation options, especially from the area of electronic institutions (Noriega 1997) and normative multi-agent systems (Boella et al. 2007). Seeking for a generic cross-disciplinary approach, we employ a formalism that

builds on Crawford and Ostrom (1995, 2005)'s *Grammar of Institutions*, borrowed from the area of institutional analysis (Ostrom 1990). The fundamental idea of the grammar is to rely on a uniform structure that allows the encoding of any form of institution (i.e., convention, norm, or rule). For this purpose, the grammar consists of an *Attributes* component (A) that describes acting individuals' characteristics, a *Deontic* component (D) used to capture the normative signal as obligation, prohibition or permission. The actual action is encoded in the *Aim* component (I), and the activation conditions (such as location, time, or previous actions) are represented in the *Conditions* component (C). Where existing, sanctions or consequences are specified in the *Or else* component (O). Using those components in varying combinations allows the capturing of different institution types. The combination of the AIC components is sufficient to express conventions (e.g., 'Drivers (A) drive (I) on the right side of the road (C).'). Social norms, in contrast, have a regulative character and include the deontic (ADIC) to describe the prohibition, permission or obligation attached to an expression (e.g., 'Drivers (A) *must* (D) drive (I) on the right side of the road (C).'). Rules, finally, exploit the entire structure (ADICO) by specifying a consequence for the expression's violation (e.g., 'Drivers (A) must (D) drive (I) on the right side of the road (C), *or else* they will be fined (O).').

While expressive in its ability to capture institutions, ADICO operates on the macro level, intended to analyse institutional outcomes in the context of institutional analysis. However, operationalising a representation that allows agents to *endogenously infer the normative function of observations at runtime* requires a refined structure, an aspect addressed by Nested ADICO (nADICO) (Frantz et al. 2013, 2015). nADICO changes the semantics for normative specifications in observations a) by allowing statements to retain information about consequences and other contextual information to substantiate the inferred understanding, and b), by allowing the combination and nesting of ADICO components to comprehensively capture actions, involved roles and actors, as well as associated normative content for both actions and consequences.

Using the rule example from above, this would translate into 'Drivers (A) must (D) drive (I) on the right side of the road (C), or else police officers (A) must (D) fine them (I) under any circumstances (C).', with the syntax `ADICADIC`. Other, more complex examples include the use of logical operators to describe the relationship between actions and consequences (e.g., `(ADIC and ADIC)ADIC` to represent the co-occurrence of actions and a single consequence; `ADIC(ADIC x/or ADIC)` to model both inclusive (`or`) and exclusive (`xor`) sanction alternatives, etc.). With those (very briefly described) refinements, this representation enables a comprehensive representation of complex behavioural traces.

The syntactic representation (*structure*) is augmented with a *process* that guides the aggregation and synthesis of observations into nADICO statements that represent an agent's normative understanding. As a first step, it involves the collection of observations under consideration of past actions, involved actors and received interaction feedback to generate institutional statements that reflect the observed behaviour. The ensuing multi-level generalisation of statements

occurs based on observable non-unique social attributes (social markers, such as roles or occupation) and combinations thereof enables the representation of subjectively generalised behaviour patterns. A detailed specification of structural aspects and the norm inference process can be found in Frantz et al. (2015); a comprehensive discussion of related literature and associated software can be found under `https://christopherfrantz.org/nested-adico`.

3.3 Intrusive vs. non-intrusive application

For its application, we differentiate between an *intrusive* and a *non-intrusive* approach, which determines the role of the discussed mechanism in the context of developed models. In both cases, agents act as observers and develop a normative understanding of the observed social and/or physical environment that is accessible to the experimenter. For the non-intrusive case, the extracted information is thus of explanatory value for the experiment observer, whereas in the intrusive case, the agent itself uses the collected information to inform its decision-making. In this case, the explanatory mechanisms thus become part of the analysed model itself.

This differentiation is essential for the flexible application of the proposed approach. While this approach is generic and "attachable" to existing agent models in the non-intrusive variant, using the proposed mechanism for agents' decision-making (i.e., feeding generalised information back into the simulation model – the intrusive application) would, of course, introduce an "ideological bias" with respect to the proposed cognitive model and associated capabilities.

4 Corruption Game

To explore this concept, we introduce an illustrative scenario that features complex interactions between different role-based actors and affords motivational autonomy of the agents based on experiential learning.

The scenario, which we refer to as the *Corruption Game*, borrows the structural characteristics of Axelrod's metanorm game (Axelrod 1986) to inform action choices, but differs in that it ignores evolutionary aspects and refines the scenario a) by explicitly modelling alternative action choices (including inaction) on the part of actors and enforcers, and b) by introducing explicit interaction of actor and second-order enforcer, aspects we will explore in detail in the following. The interaction schema of the game is depicted in Figure 1.

The narrative underlying this game is the interaction of citizens with administrative officials that may react to transgressions (e.g., corruptive behaviour) by rewarding or punishing actors. Said officials are themselves subject to oversight by second-order officials who monitor their compliance as a response to citizens' complaints. This allows the exploration of prototypical scenarios, including administrative interactions such as handling tax returns, or being punished for traffic violations, etc. – aspects that leave the first-order officials with considerable levels of discretion, making them potential participant in petty corruption.

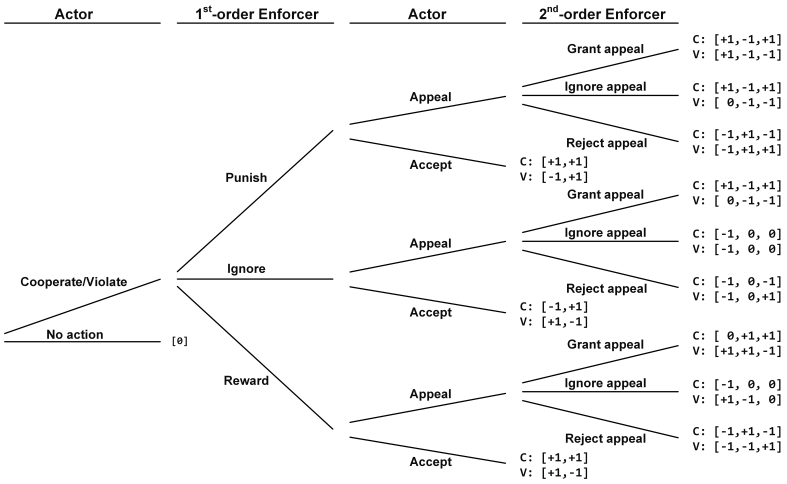

Fig. 1. Corruption Game

In the course of exploration, interesting questions revolve around the conditions under which the general behaviour shifts between violation and cooperation.

In the operationalisation, this translates into agents of two roles, either as citizens or officials (enforcers), with citizens pursuing cooperative or non-cooperative actions that are observed by enforcers, whose principal role is to reward cooperative behaviour and punish violations. As a third option, enforcers may simply ignore requests, which reflects institutional dysfunction, in contrast to wrongful decision-making by rewarding cheaters or punishing non-cheaters. Similarly, a citizen's inaction would reflect the withdrawal from economic participation. Feedback associated with enacted action-reaction combinations is applied to all involved interaction partners, and specified as part of the operationalisation.

Whatever the official's response, citizens can challenge any decision (or inaction) by appealing to a higher-level official, whose reaction determines the feedback for all involved stakeholders (citizen, first-order official, second-order official). Officials can act both as first- and second-order enforcers, but cannot act within the same transaction (i.e., an official cannot process the appeal against its own decision). The feedback for the action sequence chosen for this evaluation is denoted in Figure 1 (Syntax: `[citizen[,1stOfficial[,2ndOfficial]]]`, with `1stOfficial` and `2ndOfficial` feedback only applying where interaction with officials takes place). For this baseline exploration, the feedback structure is modelled symmetrically (i.e., the extent of negative and positive feedback is identical) and rewards correctly identified cooperative behaviour, but equally rewards undetected cheating. The reason for this largely unbiased feedback specification is the exploration of social processes that mitigate or are decisive for the convergence towards violation or cooperative behaviour. In addition to introducing a more realistic breadth of action choices, the non-binary decisions are motivated by the ability of participants to withdraw from interactions if necessary – an aspect often ignored in analytical games, but a realistic indicator to assess the impact of corruption or institutional dysfunction.

As indicated in the conceptual description in Section 3, agents develop an understanding of normative behaviour by collecting experiential observations and aggregating feedback for generalised sequences of role attributes and associated actions. For this model, we will go beyond the generalisation of observations, and allow agents to inform their action choices using such observations (intrusive approach). Agents can thus retrieve the memorised feedback in aggregated form for different initial actions (e.g., violate) to drive their decision-making.

5 Evaluation

For the evaluation of the introduced model, we parametrised the scenario with the values shown in Table 1.

Table 1. Parameters

Parameter	**Value Range and Step Size**
Number of Citizens	25 – 75; step size: 25
Number of Officials	25 – 75; step size: 25
Exploration Probability	0.1
Cheater Fraction	0.3 – 0.7; step size: 0.2
Cheating Probability	0.5 (fixed)
Weight for Observations	0.5 (fixed)
Memory Length	100 (fixed)

For the initial evaluation, we used the baseline scenario and selectively de/activated game characteristics (social learning[8], ignoring actions, appealing) and systematically varied independent variables shown in Table 1 and measured the agents' preference for cooperative (COOPERATIVE) and deviant behaviour (VIOLATE), as well as for abstinence from any interaction (INACTIVE). The condensed results are shown in the correlation overview in Table 2.[9]

The results offer a mix of expected and interesting observations. With increasing number of citizens, we observe an increase in both cooperative and violation behaviour (with a mild tendency towards violations), but more importantly, observe that actors increasingly abstain from participating in transactions. The variation of officials is likewise associated with compliance and violation, but leads to stronger levels of violation behaviour. An increasing fraction of cheating citizens leads to an overall increase in violations, which is without surprise.

Social learning in itself does not have an impact on cooperative or violation behaviour. Instead, social learning appears to lead to an overall activation of participation. If limited to specific roles (i.e., citizen, official) – tagged as 'social

[8] Social learning is operationalised as allowing agents to memorise fellow agents' institutional statement of the last action. For this operationalisation, the assumption is that all actions are overt. Agents' memory is bounded; they are able to store feedback for the last 100 experienced or observed interactions.

[9] We performed 5 runs for each parameter combination for 2000 rounds. All correlation values have been determined using Spearman's ρ.

Table 2. Correlation Overview

Parameter	COOPERATE	VIOLATE	INACTIVE
Number of Citizens	0.22	0.25	0.51
Number of Officials	0.36	0.55	0
Quota of Cheating Citizens	−0.3	0.45	0
Social Learning	−0.03	0.03	−0.25
Social Learning Separated by Role	0.32	−0.22	−0.35
Ignoring Actions	−0.38	0.36	0.51
Appealing	0.33	−0.14	−0.33

learning separated by role' –, social learning leads to stronger levels of cooperative behaviour, along with an overall stronger activation of participants.

The final two parameters, selectively preventing agents from ignoring actions and appealing, have been introduced to reduce the game in breadth (ignoring) and depth (appealing) in order to understand the effects of those actions on cooperative behaviour.[10]

At this stage, we have reviewed initial results, and little choice but to take those at face value, making the interpretation prone to the problems described in Section 2, such as the oversimplified ascription of complex behaviour to agents, and the inability to retrace the underlying processes. Following the motivation of this work, let us turn to the agents themselves and draw on their explanatory power to substantiate the insights. To achieve this, we discuss the impact of individual factors based on institutional statements recorded across all agents.

Citizen Numbers Exploring the impact of citizen numbers, the initial observation in Table 2 is the stronger engagement both on cooperative and violation side. Reviewing the relationship between number of citizens and prevalent statements en detail (see Table 3), we can make clarifying observations. Action choices resulting from an increasing number of actors have been absorbed into a few statements, here represented as simplified action sequences. These action sequences reflect generalised interaction patterns, with initiating actions noted on the right side and sequences building up to the left side. The first statement thus consists of three actions and posits that citizens accept an official's sanctioning after violating in the first place. Coming back to the specific results, in Statement 2 we can see that a side effect of the increase of citizens is an increase in mistaken rewards of violators by officials. This is insightful in that it indicates that the extent to which citizens tolerate wrongful assessment and thus institutional dysfunction.

Social Learning While our initial observations highlighted that social learning per se does neither favour cooperation nor violation, it supposedly leads to a stronger activation of participants. The statements that offer most insights (see Table 4) include the reduction in accepting an official's ignorance by cooperative citizens and reduction of citizens' inactivity, with an equal spread across other action variations (cooperation and violation).

[10] For the sake of focus, we will concentrate the discussion on the earlier parameters.

Table 3. Traces of Citizen Behaviour and Correlation with Citizen Number

Statement	Correlation
CITIZEN: ACCEPT - OFFICIAL: SANCTION - CITIZEN: VIOLATE	0.58
CITIZEN: ACCEPT - OFFICIAL: REWARD - CITIZEN: VIOLATE	0.22
CITIZEN: ACCEPT - OFFICIAL: IGNORE - CITIZEN: COOPERATE	0.3
CITIZEN: ACCEPT - OFFICIAL: REWARD - CITIZEN: COOPERATE	0.25

Table 4. Traces of Citizen Behaviour and Correlation with Social Learning

Statement	Correlation
CITIZEN: ACCEPT - OFFICIAL: IGNORE - CITIZEN: COOPERATE	−0.4
CITIZEN: IGNORE	−0.51

Social Learning separated by Role While social learning promotes unbiased participation, when looking at role-separated social learning, the results point into a different direction. In this case agents only learn from their peers, which leads to a behavioural bias towards cooperative behaviour. How does this come about?

Looking at an excerpt of the collected statements (see Table 5), we can find a clue in the faster adoption of relevant information. By learning from their peers, agents quickly adopt preferable coordination behaviour. For example, agents are quick to learn that cooperative behaviour should be rewarded (Statement 4). However, exploring statements involving appeals processes offer stronger insights into the actual dynamics. As such, officials learn to reject appeals that are lodged by violators (Statement 1), but may also quickly adopt suboptimal behaviour, such as the granting appeals to violating citizens (Statement 3), and also learn that non-reaction to appeals (Statement 2) is a potential action alternative. Without discussing all individual statements further, we can see a more refined dynamic that highlights stronger exploitation of complex institutional processes (i.e., utilising the depth of the action space).

Table 5. Traces of Citizen Behaviour and Correlation to Role-Separated Social Learning

Index	Statement	Correlation
1	OFFICIAL: REJECT_APPEAL – CITIZEN: APPEAL – OFFICIAL: SANCTION – CITIZEN: VIOLATE	0.38
2	CITIZEN: ACCEPT – OFFICIAL: IGNORE – CITIZEN: APPEAL – OFFICIAL: SANCTION – CITIZEN: VIOLATE	0.25
3	OFFICIAL: GRANT_APPEAL – CITIZEN: APPEAL – OFFICIAL: SANCTION – CITIZEN: VIOLATE	0.25
4	CITIZEN: ACCEPT – OFFICIAL: REWARD – CITIZEN: COOPERATE	0.38
5	CITIZEN: ACCEPT – OFFICIAL: IGNORE – CITIZEN: APPEAL – OFFICIAL: IGNORE – CITIZEN: COOPERATE	0.26
6	OFFICIAL: REJECT_APPEAL – CITIZEN: APPEAL – OFFICIAL: IGNORE – CITIZEN: COOPERATE	0.26
7	OFFICIAL: GRANT_APPEAL – CITIZEN: APPEAL – OFFICIAL: IGNORE – CITIZEN: APPEAL – OFFICIAL: SANCTION – CITIZEN: COOPERATE	0.13
8	OFFICIAL: REJECT_APPEAL – CITIZEN: APPEAL – OFFICIAL: SANCTION – CITIZEN: COOPERATE	0.26
9	OFFICIAL: GRANT_APPEAL – CITIZEN: APPEAL – OFFICIAL: SANCTION – CITIZEN: COOPERATE	0.23
10	CITIZEN: IGNORE	−0.35

Micro-Level Inspections While this approach allows us to retrace behavioural shifts in detail, we still operate on the macro level, based on aggregated preferred action choices of all involved agents, albeit at greater detail.

However, when exhausting the explanatory value on the macro level, the proposed approach allows us to drill deeper and explore individual agents' motivations for their behaviour. Agents develop conceptions of all explored and observed action choices, which is due to the generic nature of the approach, but it also offers a differentiated insight into the inner workings of agents, and enables us, as experimenters, to assess where cognitive processes are sufficiently represented. Figure 2 shows an extract consisting of four statements of an agent's runtime understanding (in original syntax), centred around its decision-making with respect to the action 'appeal'. The statements clearly show how agents can operate with conflicting signals. The first statement, for example, suggests that an agent should appeal (positive deontic value) after an official's punishment of its violation. The motivation for this is reduced to the observation that chances are that the official may actually grant the appeal. The second statement effectively explores the opposing signal of discouraging appealing because of potential rejection. Similarly, the last two statements highlight conflicting motivations as to whether appealing after showing cooperative behaviour is useful.

```
A=A(*, {ROLE=[CITIZEN]}), D=3.0, I=I(APPEAL, *), C=C({PREVIOUS_ACTION=L0: A=A(*, {ROLE=[OFFICIAL]}), I=I(SANCTION, *),
        C=C({PREVIOUS_ACTION=L0: A=A(*, {ROLE=[CITIZEN]}), I=I(VIOLATE, *), C=C(*), O=(null)}), O=(null)}),
  O=(L1: A=A(*, {ROLE=[OFFICIAL]}), D=-3.0 (inv), I=I(GRANT_APPEAL, *), C=C(*), O=(null))

A=A(*, {ROLE=[CITIZEN]}), D=-1.0, I=I(APPEAL, *), C=C({PREVIOUS_ACTION=L0: A=A(*, {ROLE=[OFFICIAL]}), I=I(SANCTION, *),
        C=C({PREVIOUS_ACTION=L0: A=A(*, {ROLE=[CITIZEN]}), I=I(VIOLATE, *), C=C(*), O=(null)}), O=(null)}),
  O=(L1: A=A(*, {ROLE=[OFFICIAL]}), D=1.0 (inv), I=I(REJECT_APPEAL, *), C=C(*), O=(null))

A=A(*, {ROLE=[CITIZEN]}), D=-0.5, I=I(APPEAL, *), C=C({PREVIOUS_ACTION=L0: A=A(*, {ROLE=[OFFICIAL]}), I=I(REWARD, *),
        C=C({PREVIOUS_ACTION=L0: A=A(*, {ROLE=[CITIZEN]}), I=I(VIOLATE, *), C=C(*), O=(null)}), O=(null)}),
  O=(L1: A=A(*, {ROLE=[OFFICIAL]}), D=0.5 (inv), I=I(REJECT_APPEAL, *), C=C(*), O=(null))

A=A(*, {ROLE=[CITIZEN]}), D=0.5, I=I(APPEAL, *), C=C({PREVIOUS_ACTION=L0: A=A(*, {ROLE=[OFFICIAL]}), I=I(REWARD, *),
        C=C({PREVIOUS_ACTION=L0: A=A(*, {ROLE=[CITIZEN]}), I=I(COOPERATE, *), C=C(*), O=(null)}), O=(null)}),
  O=(L1: A=A(*, {ROLE=[OFFICIAL]}), D=-0.5 (inv), I=I(GRANT_APPEAL, *), C=C(*), O=(null))
```

Fig. 2. Excerpt of Micro-level Statements for Action 'APPEAL'

Following this brief exposition, we can see that the mechanism is not only able to leverage explanatory insights into behavioural changes on the macro level, but also to transparently represent individual-level cognitive processes, such as the notion of cognitive dissonance (Festinger 1957) as just discussed above.

6 Discussion

In this paper, we argued for the use of a generic agent conception that satisfies a fundamental subset of processes found in social animals (stereotyping, social learning), and specifically humans, and attach or integrate this mechanism with existing agent-based models, so as to leverage these processes to provide additional explanatory power – in addition to the conventional aggregate macro-level observation of dependent variables. To explore model internals, agents collect

and generalise their observations, and represent those in a uniform way that allows their aggregation on arbitrary level of social organisation (e.g., individual observations, groups, or society at large). We showcased this approach using a moderately complex institutional scenario in order to explore the emerging behaviour at greater depth.

This work intends to drive the discussion around *exploring agency to understand agency* using institutional mechanisms. Given this motivation, the approach presented here is a candidate operationalisation in the form of a domain-independent baseline architecture that allows for the consideration of fundamental functions of human operation in social environments. Questions that invite for further discussion revolve around the minimal cognitive functions sufficient for a baseline operationalisation (here: stereotyping), and, if used as input for decision-making, in how far such architecture affords an 'ideological buy in' by experimenters and affects modelling freedom.

But returning to the motivation of this work, what are the concrete benefits of shifting from descriptive macro-level to explanatory micro-level approaches?

- Agents can be used as passive, non-intrusive observers, e.g., only used for verification of the model, or for the inspection of specific runs. The condensed generative conception of the institutional environment can thus be used for methodological support during model development.
- Using a generic institution operationalisation allows the detection of both intentional and unintentional behaviour (independent of the non-/intrusive application). This provides the experimenter with insight into both 'desirable' and 'undesirable' behaviours that may withdraw themselves from experimental observation for cases in which the underlying dynamics are obscured by aggregate metrics. This aspect is of analytical value, since it allows the experimenter to retrace explicit explanatory links between micro-level interaction dynamics and macro-level phenomena.

Both such aspects have the potential of contributing to building greater confidence in the development and analysis of agent-based models, and consequently, for the application of agent-based modelling more generally.

Future efforts involve the application of this approach to existing datasets to explore its usability for real-world applications, both in terms of usefulness and efficiency. This includes the intent to make this mechanism more readily available, e.g., as a plugin, to explore its value with new or existing simulation scenarios. Reflecting on the choice of an institution representation from the area of institutional analysis opens up manifold further interdisciplinary application opportunities. However, whatever the chosen mechanism, the essential argument this paper makes is that any approach to develop explanatory ABMs will, in one way or another, have to consider institutional concepts at its basis.

Concluding, we believe that it is important to drive a cross-disciplinary consensus that agent-based modelling has the capability to explore complex social phenomena, but unlike other quantitative approaches, can also offer ways to facilitate the interpretation of its own operation – and which vehicle would be more self-referential than the agents themselves?

Bibliography

Abar, S., Theodoropoulos, G. K., Lemarinier, P., and OHare, G. M. (2017). Agent based modelling and simulation tools: A review of the state-of-art software. *Computer Science Review*, 24:13 – 33.

Axelrod, R. (1986). An Evolutionary Approach to Norms. *The American Political Science Review*, 80(4):1095–1111.

Balke, T. and Gilbert, N. (2014). How Do Agents Make Decisions? A Survey. *Journal of Artificial Societies and Social Simulation*, 17(4):13.

Bandura, A. (1977). *Social Learning Theory*. General Learning Press, New York (NY).

Bianchi, F. and Squazzoni, F. (2015). Agent-based models in sociology. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(4):284–306.

Birks, D., Townsley, M., and Stewart, A. (2012). Generative explanations of crime: Using simulation to test criminological theory. *Criminology*, 50(1):221–254.

Boella, G., van der Torre, L., and Verhagen, H. (2007). Introduction to Normative Multi-Agent Systems. In Boella, G., van der Torre, L., and Verhagen, H., editors, *Normative Multi-agent Systems, Dagstuhl Seminar Proceedings 07122*, Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

Castelfranchi, C. (2014). Minds as social institutions. *Phenomenology and the Cognitive Sciences*, 13(1):121–143.

Cederman, L. (2005). Computational models of social forms: Advancing generative process theory. *American Journal of Sociology*, 110(4):864–893.

Crawford, S. E. and Ostrom, E. (1995). A Grammar of Institutions. *The American Political Science Review*, 89(3):582–600.

Crawford, S. E. and Ostrom, E. (2005). A Grammar of Institutions. In *Understanding Institutional Diversity*, chapter 5, pages 137–174. Princeton University Press, Princeton (NJ).

Edmonds, B. and Moss, S. (2005). From KISS to KIDS – An 'Anti-simplistic' Modelling Approach. In Davidsson, P., Logan, B., and Takadama, K., editors, *Multi-Agent and Multi-Agent-Based Simulation*, volume 3415 of *Lecture Notes in Computer Science*, pages 130–144. Springer, Berlin.

Elsenbroich, C. (2012). Explanation in agent-based modelling: Functions, causality or mechanisms? *Journal of Artificial Societies and Social Simulation*, 15(3):1.

Epstein, J. M. (1999). Agent-Based Computational Models and Generative Social Science. *Complexity*, 4:41–60.

Farmer, J. D. and Foley, D. (2009). The economy needs agent-based modelling. *Nature*, 460:685–686.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press, Palo Alto, CA.

Frantz, C., Purvis, M. K., and Nowostawski, M. (2014). Agent-Based Modeling of Information Transmission in Early Historic Trading. *Social Science Computer Review*, 32(3):393–416.

Frantz, C., Purvis, M. K., Nowostawski, M., and Savarimuthu, B. T. R. (2013). nADICO: A Nested Grammar of Institutions. In Boella, G., Elkind, E., Savarimuthu, B. T. R., Dignum, F., and Purvis, M. K., editors, *PRIMA 2013: Principles and Practice of Multi-Agent Systems*, volume 8291 of *Lecture Notes in Artificial Intelligence*, pages 429–436, Berlin. Springer.

Frantz, C. K., Purvis, M. K., Savarimuthu, B. T. R., and Nowostawski, M. (2015). Modelling Dynamic Normative Understanding in Agent Societies. *Scalable Computing: Practice and Experience*, 16(4):355–378.

Galán, J. M. and Izquierdo, L. R. (2005). Appearances can be deceiving: lessons learned re-implementing Axelrod's 'evolutionary approach to norms'. *Journal of Artificial Societies and Social Simulation*, 8(3):2.

Galán, J. M., Izquierdo, L. R., Izquierdo, S. S., Santos, J. I., del Olmo, R., and López-Paredes, A. (2017). *Checking Simulations: Detecting and Avoiding Errors and Artefacts*, pages 119–140. Springer International Publishing, Cham.

Gilbert, N. (2008). *Agent-Based Models*. Sage Publications, London.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., and Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109(1):3–25.

Grüne-Yanoff, T. (2009). The explanatory potential of artificial societies. *Synthese*, 169(3):539–555.

Hodgson, G. M. (2006). What Are Institutions? *Journal of Economic Issues*, 40(1):1–25.

Jackson, J. C., Rand, D., Lewis, K., Norton, M. I., and Gray, K. (2017). Agent-based modeling: A guide for social psychologists. *Social Psychological and Personality Science*, 8(4):387–395.

Jager, W. (2017). Enhancing the realism of simulation (eros): On implementing and developing psychological theory in social simulation. *Journal of Artificial Societies and Social Simulation*, 20(3):14.

Janssen, M. and Jager, W. (1999). An integrated approach to simulating behavioural processes: A case study of the lock-in of consumption patterns. *Journal of Artificial Societies and Social Simulation*, 2(2).

Kahneman, D. (2013). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York (NY).

Kravari, K. and Bassiliades, N. (2015). A survey of agent platforms. *Journal of Artificial Societies and Social Simulation*, 18(1):11.

Lengnick, M. (2013). Agent-based macroeconomics: A baseline model. *Journal of Economic Behavior & Organization*, 86:102 – 120.

Levy, S., Martens, K., and van der Heijden, R. (2016). Agent-based models and self-organisation: addressing common criticisms and the role of agent-based modelling in urban planning. *Town Planning Review*, 87(3):321–338.

Luke, S., Cioffi-Revilla, C., Panait, L., Sullivan, K., and Balan, G. (2005). MASON: A Multiagent Simulation Environment. *Simulation*, 81(7):517–527.

Macy, M. and Tsvetkova, M. (2015). The signal importance of noise. *Sociological Methods & Research*, 44(2):306–328.

Minar, N., Burkhart, R., Langton, C., and et al. (1996). The swarm simulation system: A toolkit for building multi-agent simulations. Technical report.

Müller, B., Bohn, F., Dreler, G., Groeneveld, J., Klassert, C., Martin, R., Schlter, M., Schulze, J., Weise, H., and Schwarz, N. (2013). Describing human decisions in agent-based models odd+d, an extension of the odd protocol. *Environmental Modelling & Software*, 48:37 – 48.

Noriega, P. (1997). *Agent-Mediated Auctions: The Fishmarket Metaphor*. PhD thesis, Universitat Autonoma de Barcelona, Barcelona.

North, D. C. (1990). *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, New York (NY).

North, D. C. (1991). Institutions. *Journal of Economic Perspectives*, 5(1):97–112.

North, M. J., Collier, N. T., Ozik, J., Tatara, E., Altaweel, M., Macal, C. M., Bragen, M., and Sydelko, P. (2013). Complex adaptive systems modeling with Repast Simphony. *Complex Adaptive Systems Modeling*, 1(1):3.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, New York (NY).

Pan, X., Han, C. S., Dauber, K., and Law, K. H. (2007). A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. *AI & Society*, 22(2):113–132.

Peirce, C. S. (1878). How to Make Our Ideas Clear. *Popular Science Monthly*, 12:286–302.

Schelling, T. C. (1971). Dynamic Models of Segregation. *The Journal of Mathematical Sociology*, 1(2):143–186.

Searle, J. R. (2005). What is an Institution? *Journal of Institutional Economics*, 1(1):1–22.

Shults, F. L., Gore, R., Wildman, W. J., Lynch, C., Lane, J. E., and Toft, M. (2017). Mutually escalating religious violence: A generative model. In *Proceedings of the 2017 Social Simulation Conference, Dublin, Ireland, 2017*.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Tisue, S. and Wilensky, U. (2004). NetLogo: Design and Implementation of a Multi-Agent Modeling Environment. In *SwarmFest*, Ann Arbor (MI).

Tolk, A. (2015). *Learning Something Right from Models That Are Wrong: Epistemology of Simulation*, pages 87–106. Springer International Publishing, Cham.

Zeithamova, D., Schlichting, M., and Preston, A. (2012). The hippocampus and inferential reasoning: building memories to navigate future decisions. *Frontiers in Human Neuroscience*, 6:70.