Superstition in the Cognitive Model: Modelling Ritualised Behaviour as Error Management

Mariusz Rybnik, Ivan Puga-Gonzalez, F. LeRon Shults, Paweł Łowicki, Ewa Dąbrowska-Prokopowska, Andrew Atkinson, and Konrad Talmont-Kaminski

Abstract Skinner found that superstitious behaviour in pigeons results from accidental operant conditioning. We use a simple cognitive model based upon reinforcement learning to show that ritualisation of behaviour arises in analogous conditions. This makes it possible to model the creation of ritual traditions using minimal means in agent-based models, thereby opening a novel and potentially highly fruitful approach to the study of this highly significant human behaviour.

Keywords Error management · Cognitive model · Superstition · Ritualised behaviour

1 Introduction

Most research into ritualised behaviour tends towards high-level explanations that involve, among other things, the cultural transmission of supernatural beliefs. The explanation for the spontaneous appearance of superstitious behaviours in pigeons provided by the behaviourist psychologist B.F. Skinner could not be any more different [9]. It sought to explain superstitious behaviour merely by reference to acci-

The research leading to these results has received funding from the Norwegian Financial Mechanism 2014–2021 (Project number 2019/34/H/HS1/00654).

M. Rybnik (✉)
Faculty of Computer Science, University of Bialystok, Bialystok, Poland
e-mail: m.rybnik@uwb.edu.pl

I. Puga-Gonzalez · F. L. Shults
NORCE Center for Modeling Social Systems, Kristiansand, Norway

P. Łowicki
Faculty of Psychology, University of Warsaw, Warsaw, Poland

E. Dąbrowska-Prokopowska
Institute of Sociology, University of Bialystok, Bialystok, Poland

A. Atkinson · K. Talmont-Kaminski
Society and Cognition Unit, University of Bialystok, Bialystok, Poland

C. Elsenbroich and H. Verhagen (eds.), *Advances in Social Simulation*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-031-57785-7_44

dental operant conditioning. Understood as showing that superstitious behaviour arises spontaneously where an agent—artificial or not—attempt to control an environment that is unpredictable, this line of research has been expanded upon greatly by researchers such as Ono, who showed the same effect in humans [8], as well as as Killeen, who showed that pigeons placed in analogous conditions vary their rate of superstitious behaviour in a way that maximises their pay-off given uncertainty [7]. Killeen's observation has been generalised by Haselton and Buss, who showed that human and other evolved cognitive mechanisms will be biased in favour of committing the least costly errors when operating under uncertainty, i.e., they will exhibit error management [5]. Given the need to detect potential threats/opportunities in the environment, people will avoid false negative errors at the cost of making more false positive ones. This includes overdetecting causal connections and agency. Further research has expanded the scope of this type of explanation even more by showing that the kind of behaviour Skinner observed is a normal by-product even for formal or computing mechanisms that attempt to predict system behaviour under conditions that are often met with in natural settings [2, 4]. In effect, researchers working in the Skinner tradition have shown that superstitious behaviour of all agents—be they artificial or natural—is to be explained as a by-product of the epistemic circumstances encountered when attempting to control any unpredictable system.

The aim of the line of research pursued here is to see whether ritualised behaviour can be explained in the same general terms, i.e., as the product of false positive errors resulting from error management, and to explore how changes in conditions affect spontaneous ritualisation in silico. Our ultimate goal is to create an agent-based model, to be used to study the emergence of stable rituals at the group level. In other words, we are seeking to show that rituals, including group rituals, are also to be ultimately explained in terms of a fundamental phenomenon that affects all agents given the right set of basic conditions—with most aspects of rituals as normally discussed (including supernatural beliefs or the role of anxiety) being ancillary (see [10]). The definition of ritualised behaviour used by us is drawn from Boyer and Lienard [3] where they identify five basic characteristics common to ritualised behaviours. Three are investigated here: redundancy—including elements in one's behaviour that are not necessary to achieve the apparent aim; goal-demotion—including elements in one's behaviour that have no apparent effect; rigidity—performing all of the elements, including the redundant and goal-demoted ones, in a strict order. The modelled behaviour is studied with the use of the PathGame—a (pseudo)game with clear and easily manipulable rules that makes operationalisation and analysis straight forward.

2 The PathGame

PathGame is a novel methodology developed in order to empirically investigate the conditions under which humans, as well as computational cognitive models, spontaneously ritualise their behaviour. It is based upon Stuart Vyse's earlier research [6, 12]

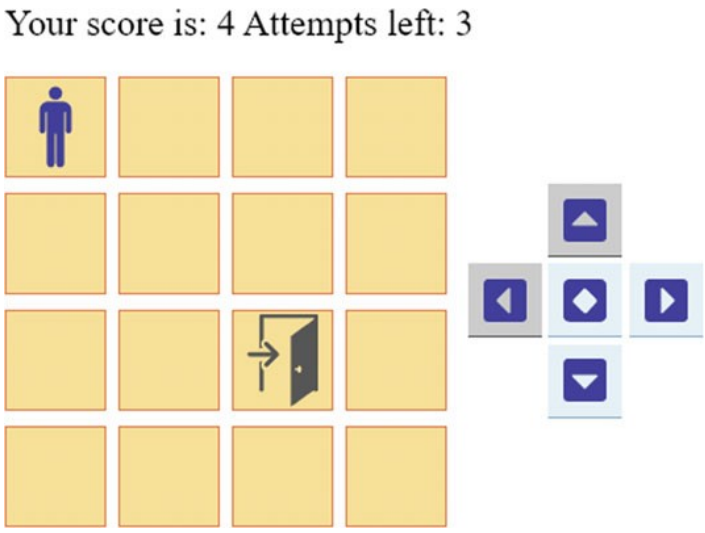


Fig. 1 PathGame implemented as a web application using TypeScript

into superstitious behaviour. Vyse's work was explicitly pursued in the Skinnerian tradition described above, so expanding upon helps to make the connection between superstitious and ritualised behaviours explicit. Getting humans and artificial agents to deal with the same problem allows direct comparison of their behaviours.

The game takes place on a four-by-four matrix. The basic goal is to move the avatar from the starting position in the top-left corner to the exit situated two blocks down and two to the right. Five buttons are available: four directional ones and a fifth, pressing which does not lead to any apparent effect. Players are allowed to play the game fifty times, sometimes receiving points upon reaching the exit cell, with the aim to get the maximum number of points. The interactive game's interface (implemented as a TypeScript web application for the purpose of human studies; not presented here) is shown in Fig. 1. **Redundancy** is operationalised as pressing the up or left buttons—pressing them is unnecessary to reach the exit. **Goal-demotion** is operationalised as pressing the fifth 'mystery' button—pressing it has no effect, apart from being recorded. **Rigidity** is operationalised as repeating (within ten attempts) the same non-minimal path, i.e., one that includes redundancy or goal-demotion. Rigidity is distinguished from **automaticity**, which is operationalised as repeating any minimal path within ten attempts, and which is understood to result in humans from minimisation of cognitive effort rather than ritualisation.

When faced with Vyse's simpler set-up (only the right and down buttons), human players tended to correctly represent the game's winning conditions when they were predictable but generated complex and completely fictitious hypotheses when points were awarded randomly. In our own pilot studies with human subjects, we found that redundancy, goal-demotion and rigidity all increased when the probability of obtaining points at each attempt was low, but automaticity grew as success was

more likely. In effect, people appeared to be learning to form incorrect associations between the ritualised aspects of their behaviour and obtaining points when points were obtained randomly and rarely. As Vyse noted, "operant conditioning is not just for rats and pigeons" [11].

3 The Cognitive Model (CM)

In order to provide an artificial model of spontaneous ritualisation using the *PathGame* paradigm, a computational CM was implemented using the Anylogic environment [1]. The goal was to determine whether spontaneous ritualisation of behaviour qualitatively akin to that exhibited by humans could be generated using a simple artificial system designed to identify non-random patterns, thereby showing a common underlying basis for the phenomenon. The CM works on the basis of a set of weights on each of the cells in the matrix, which determine the probability that the CM will make a particular move when the avatar is in that cell. The avatar moves around the matrix, with the each move determined stochastically on the basis of the weights at the currently-occupied cell, until the exit cell is reached. Depending upon whether a point was obtained at the end of the walk, the weights of the moves made are either weakened or reinforced.

3.1 Terms and Parameters

A *Move* is regarded here as a movement to another valid cell or the *Goal-Demotion* pseudo-activity. *Moves* are denoted as the four cardinal geographical directions on a map: N, E, S, W, with Goal-Demotion denoted as the question mark: ? A *Walk* is defined as a sequence of *Moves* leading from the starting position to the exit cell and is represented symbolically as a *Path*, such as: $SSEW?EE$. A single *Game* consists of a given number of *Walks*, which for the purpose of this research was always 50. CM try to gain and store the knowledge about rewards along the *Game*.

The game matrix with initial and final move weights is presented in Fig. 2. A *Move weight* is a floating-point number associated with the specific *move* at the specific cell. The *move weights* are depicted in each cell (with the obvious exception of the exit cell marked as a circle). The numbers indicate weights for each direction, with the number in the center being the *Goal-Demotion* weight. The *Move probability* of move X at a cell (i, j) can be obtained by dividing the corresponding weight t_{Xij} by the sum of all weights $\sum t_{xij}$, where $x \in \{N, E, S, W, ?\}$. Initial weights are set to ensure the CM-controlled avatar is likely to reach the exit cell in a small number of moves (just as is the case with human players). Reinforcement/attenuation during the game changes these weights significantly leading to very different behaviour depending upon the reinforcement schedule.

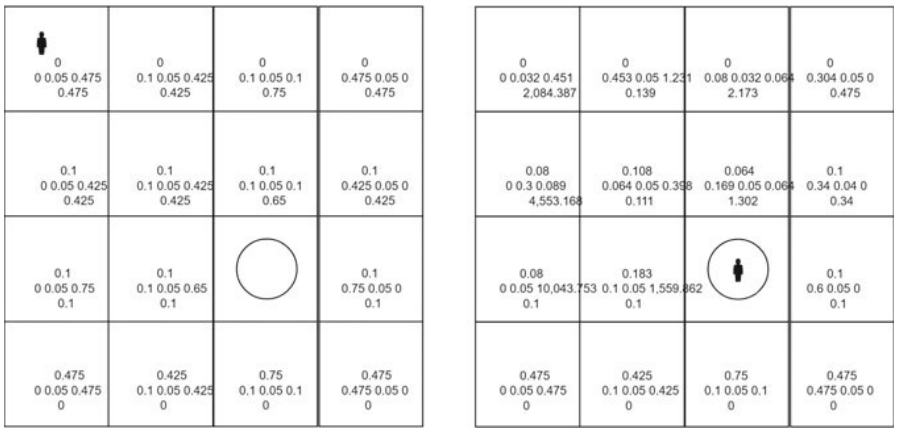


Fig. 2 The game matrix with starting move weights (on the left), and the game matrix with final move weights (on the right)

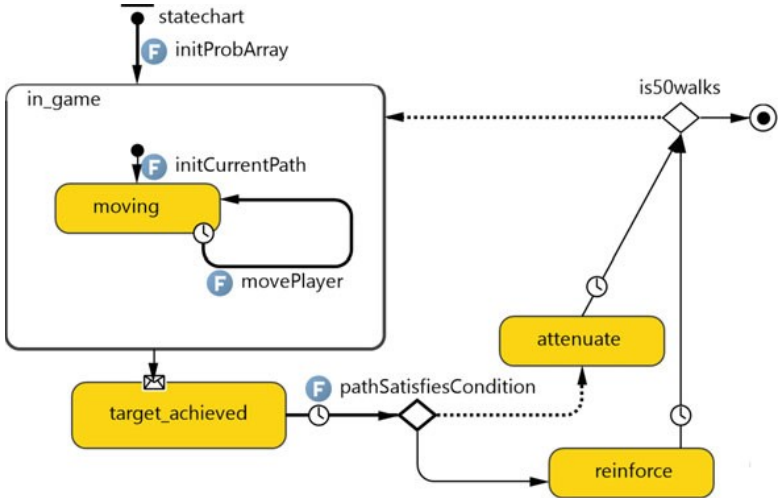


Fig. 3 State diagram of CM (in Anylogic)

3.2 State Diagram

The CM completes 50 *walks*. After each *walk*, *Reinforcement Learning* or *Attenuate Learning* is performed according to a random or non-random schedule. The state diagram of the CM implemented in Anylogic is presented in Fig. 3.

During the game the CM performs the following steps:

Step 1. Initialise move weights
Step 2. Make random weight-determined moves till end cell reached

Step 3. Schedule determines whether to reinforce/attenuate the walk
Step 4A. Reinforcement learning
OR
Step 4B. Attenuate learning
Step 5. While less than 50 walks loop back to step 2.

3.3 Reinforcement Versus Attenuation

If a *walk* is rewarded, reinforcement occurs—the *moves* in the *walk* have their *move weights* increased as follows: $t_{Xij} \rightarrow t_{Xij} \times Reinforce\,Ratio$, where $Reinforce\,Ratio > 1$. In effect, future *walks* are more likely to include the same *moves*. Due to stochastic nature of the CM however, some degree of indeterminacy is maintained, thereby allowing experimentation.

If a *walk* is not rewarded, attenuation occurs—the *move weights* contained in the specific *path* are attenuated: $t_{Xij} \rightarrow t_{Xij} \times Attenuate\,Ratio$, where $Attenuate\,Ratio \in (0; 1)$. In effect, in future future *walks*, the *CM* is less likely to perform the same *moves*, as they were not beneficial.

In addition, to avoid the CM generating very long paths, path length mitigation is introduced by making reinforcement/attenuation proportional to path length, with the minimal path length of 4 being regarded as the base value.

4 Testing Methodology

Given that on the Skinnerian approach being pursued here, ritualised behaviour is understood as arising due to accidental operant conditioning, it is important to test the CM in two different kinds of scenarios—random and non-random. In the non-random scenarios, the CM should be able to identify the winning paths while in the random scenarios it should generate ritualised behaviour.

4.1 Non-random Scenarios

The non-random scenarios are based on a set of predefined, alternative rules that reward specific CM behaviour. These rules are based on cells the path has to go through or avoid or specific buttons being pressed or not pressed. To test the CM in more complex non-random scenarios, some include pairs of these rules.

In these scenarios, a path increases the score by one and is reinforced if it satisfies the rule in force in the game being played. Otherwise, the path is attenuated and the score remains unchanged. Where a pair of rules is in force, both have to be satisfied by a path for it to be reinforced and for the score to increase by one.

Basic Rules

– Cell-Based Rules

R1. Avoid the bottom left quadrant
R2. Avoid the top right quadrant
R3. Walk through any 4th row cell
R4. Walk through any 4th column cell

– Button-Based Rules

R5. Press the up arrow at least once
R6. Press the left arrow at least once
R7. Never press the goal-demotion button
R8. Press the goal-demotion button at least once
R9. Press the goal-demotion button at least twice.

Basic Rule Pairs

– **R1 and R3**: Avoid the bottom left quadrant and Walk through any 4th row cell
– **R1 and R5**: Avoid the bottom left quadrant and Press the up arrow at least once
– **R1 and R7**: Avoid the bottom left quadrant and Never press the goal-demotion button
– **R2 and R4**: Avoid the top right quadrant and Walk through any 4th column cell
– **R2 and R6**: Avoid the top right quadrant and Press the up arrow at least once
– **R2 and R7**: Avoid the top right quadrant and Never press the goal-demotion button
– **R3 and R4**: Walk through any 4th row cell and through any 4th column cell
– **R3 and R6**: Walk through any 4th row cell and Press the left arrow at least once
– **R3 and R7**: Walk through any 4th row cell and Never press the goal-demotion button
– **R4 and R5**: Walk through any 4th column cell and Press the up arrow at least once
– **R4 and R7**: Walk through any 4th column cell and Never press the goal-demotion button
– **R5 and R7**: Press the up arrow at least once and Never press the goal-demotion button
– **R6 and R7**: Press the left arrow at least once and Never press the goal-demotion button.

4.2 Random Scenarios

In these scenarios, reward and the associated reinforcement/attenuation schedule are independent of the path used and depend merely upon the *Reinforce Probability* parameter, which gives the stochastic probability of reward/reinforcement.

5 Results

5.1 Results for Non-random Scenarios

The goal of these scenarios was to test whether and to what degree the CM was able to extract rules knowledge from the received rewards and to then use that knowledge to increase the number of points obtained over the length of the game. As such, the scores obtained by the CM have to be compared to the scores obtained without any reinforcement—these show how difficult the rules are to satisfy purely randomly and differ between the different non-random rules tested.

The scores obtained by the CM for the basic rules and their pairs are presented in Fig. 4. These are the mean scores for 1000 *games* of 50 walks. Each *game* is initialized in the same way (including the same initial move weights) and with the following parameter values:

- *Reinforce Ratio* = 9
- *Attenuate Ratio* = 0.8
- *Path Size Mitigation* applied linearly (see Sect. 3.3 for details).

The rules tested proved to vary greatly in terms of how easy they were to happen upon randomly, with the rule pairs generally proving more difficult. This pattern also held for cases where reinforcement was in place. However, in every case tested, reinforcement allowed the CM to obtain a clearly increased average score. In fact, it was with rule pairs that the effect of reinforcement was the most striking. Of course, the difficulty of individual rules could be modified by altering the initial weights—for example, pressing the goal-demotion button could be made much more likely, thereby making R9 much more likely to be satisfied randomly. So, not much can be read into individual rules. However, it is the overall pattern of reinforcement increasing scores that shows that for the set of rules we tested, the reinforcement learning was successful and that this CM is a satisfactory model of pattern-seeking behaviour. Having observed this behaviour, we could go on to test how this CM behaved when presented not with predictable scenarios but with purely stochastic ones.

5.2 Results for Random Scenarios

Having established that the CM is capable of learning non-random rules, it was necessary to check whether this was sufficient for the CM to generate ritualised behaviour when presented with random scenarios, as operationalised in terms of redundancy, goal-demotion and rigidity (see Sect. 2).

CM configuration is as follows:

- parameter *Reinforcement Probability* $\in < 0 : 0.01 : 1 > -101$ values

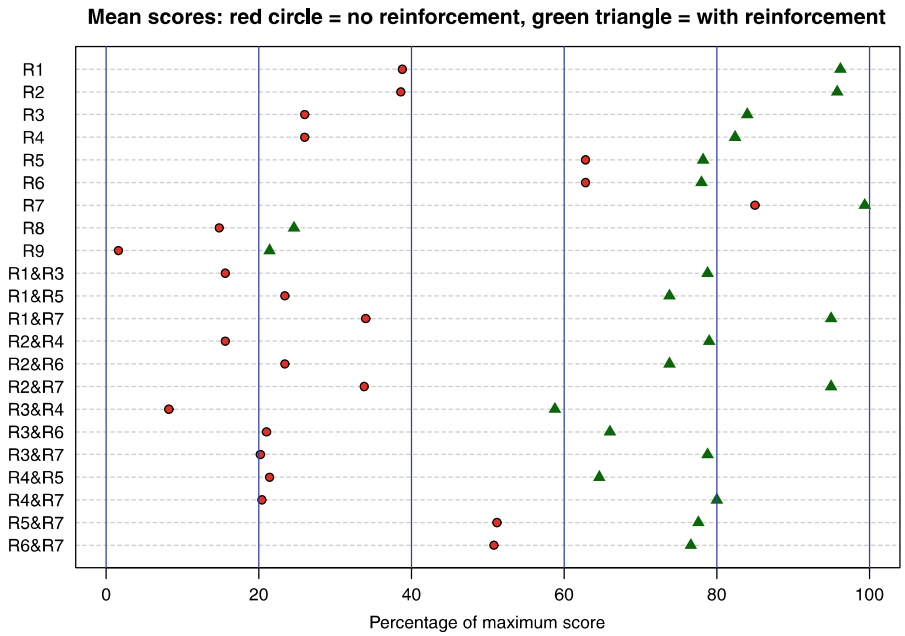


Fig. 4 CM mean scores for the individual rules and selected rule pairs, comparing reinforcement with no reinforcement

– 1000 *Games* per each *Reinforcement Probability* value.
– *Reinforce Ratio* = 9
– *Attenuate Ratio* = 0.8
– *Path Size Mitigation* applied linearly (see Sect. 3.3 for details).

A total of 101,000 *Games* were run. Each simulation consisted of 50 time steps (i.e., 50 *Walks* or *Paths*). Rewards were awarded stochastically based on the value of *Reinforcement Probability* and independently of the path used.

The metrics for the four characteristics obtained from the experiments are presented in Fig. 5.

Goal-Demotion—as shown in the upper left plot in Fig. 5—occurs in the majority of paths when reinforcement is rare but drops away and is almost never seen in cases where the probability of reinforcement is above 0.3. This is in line with human behaviour, where it has been observed that goal-demotion also is much more common when paths are rarely reinforced.

Redundancy—as shown in the upper right plot in Fig. 5—is very high at low *Reinforcement Probability* values—occurring in the majority of walks—but then drops towards zero as the probability of reward increases, similarly to goal-demotion. This pattern has also been observed in human players.

Automaticity—shown in the bottom left plot in Fig. 5—is not a necessary characteristic of ritualised behaviour but was of interest to us given the methodology used. It behaved in the opposite manner to goal-demotion and redundancy, in that it was

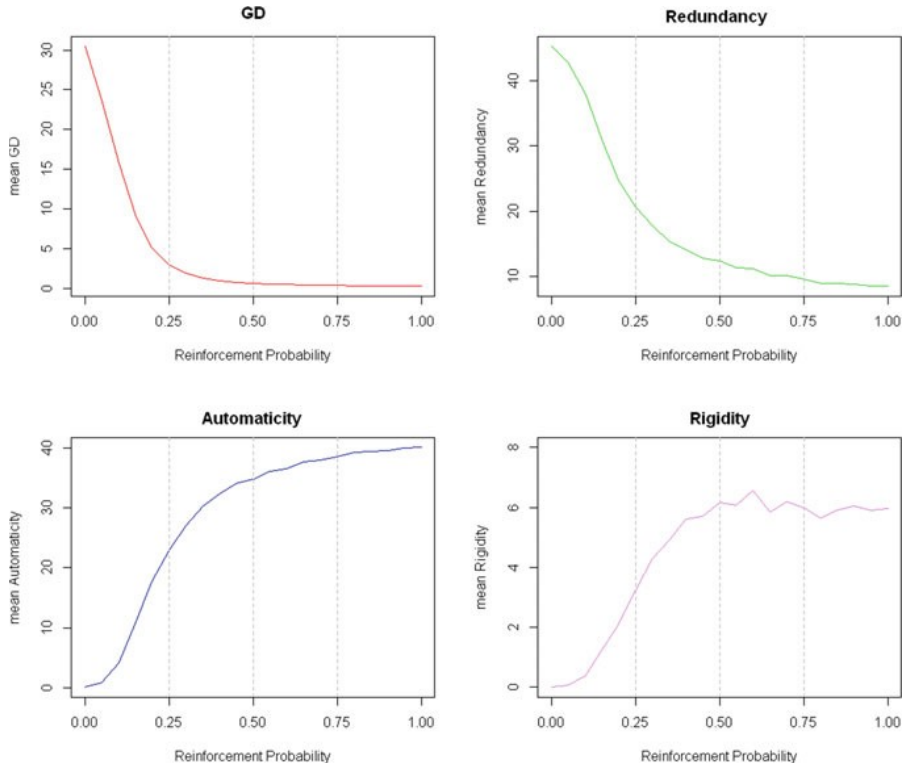


Fig. 5 Average error management metrics versus *Reinforcement Probability*

almost non-existent when reinforcement was low and quickly came to dominate at higher reinforcement probabilities. This pattern is also analogous to the one observed with human players.

Rigidity—shown in the bottom right plot in Fig. 5—proved to be the problematic measure. As can be seen, it is quite rare at high *Reinforcement Probability* values but is completely absent when reinforcement drops to zero. As such it will be discussed at some length in the final section. Rigidity is regarded here as the repetition of a non-minimal path within the ten most recent walks. It is low with low *Reinforcement Probability* values (bottom right plot in Fig. 5). Due to rare learning the paths tend to be quite long and very stochastic, so they rarely repeat subsequently at all. Rigidity rises quickly with the increase of *Reinforcement Probability* and stabilizes obtaining about 6–7 occurrences out of 50 with *Reinforcement Probability* $\in (0.25, 1)$. It could be interpreted that the paths tend to be more focused and repetitive, both minimal ones that conform to *Automaticity* and non-minimal ones that conform to *Rigidity*. It is also worth noting that *Rigidity* drops slightly within the mentioned range, probably due to *Automaticity* raising, that overtake also non-minimal repetitions.

6 Discussion

The cognitive model was able to learn simple rules in the non-random scenarios, thereby showing that we developed a simple learning model that could be usefully compared to the behaviour generated by humans. When presented with random scenarios, it spontaneously generated redundant and goal-demoted behaviour characteristic of rituralised behaviour, with both redundancy and goal-demotion occurring more commonly in conditions that also favour that behaviour in humans. Rigidity proved more elusive. While random scenarios did indeed generate rigidity, they did not do so in the pattern met with in humans, where rigidity tends to co-vary with goal-demotion and redundancy. However, this difference in results is to be understood in terms of the limitations of the learning model used rather than as a problem with the main hypothesis. The reason is that, when tested, humans tend to rigidify their behaviour in a manner that was not open to the CM tested here. Specifically, they tend to form series of paths that they then use repeatedly, believing that the rules determining success change from walk to walk in a set pattern. In effect, they rigidify not over individual paths but over sets of them. This was not possible for the CM we created, as it has no memory of paths used as its behaviour is merely determined on the basis of the move weights.

Keeping in mind this significant limitation, it is possible to conclude that the study has been successful in showing that ritualisation of behaviour, or at least some aspects of it, is simply caused by accidental operant conditioning working in a scenario that is unpredictable—in line with the explanation that Skinner gave for superstition. As such, superstition and ritual must be seen as having at least in part a common epistemic basis that means that we should expect such behaviour in any system capable of learning when it is placed in the relevant set of conditions. This means in particular, that supernatural beliefs that commonly are connected to rituals are not basic to them, nor is anxiety necessary to generate ritualised behaviour.

The behaviour of the CM is also interesting in view of error management theory. When dealing with relatively easy random scenarios in which success was highly likely, the CM rapidly ended up automatising its behaviour while low success rates led to much greater innovation. To fully explore this aspect of ritualisation, however, it would be necessary to make each move costly thereby putting the CM in a position where it has to decide whether to use a more costly non-minimal path that may in some scenarios be required to earn points.

In future modelling work we will explore spontaneous ritualisation of behaviour using more complex cognitive models capable of rigidifying over sets of paths and then create an agent-based model to explore the social conditions necessary to sustain a ritual-tradition.

References

1. AnyLogic: Simulation Modeling Software Tools and Solutions (2023). https://www.anylogic.com/
2. Beck, J., Forstmeier, W.: Superstition and belief as inevitable by-products of an adaptive learning strategy. Hum. Nat. **18**, 35–46 (2007)
3. Boyer, P., Liénard, P.: Why ritualized behavior? Precaution Systems and action parsing in developmental, pathological and cultural rituals. Behav. Brain Sci. **29**(6), 595–613 (2006)
4. Foster, K.L., Kokko, H.: The evolution of superstitious and superstition-like behavior. Proc. R. Soc. B: Biol. Sci. **276**, 31–37 (2009)
5. Haselton, M., Buss, D.: Error management theory: a new perspective on biases in cross-sex mind reading. J. Pers. Soc. Psychol. **78**, 81–91 (2000)
6. Heltzer, R., Vyse, S.: Intermittent consequences and problem solving: the experimental control of "superstitious" beliefs. Psychol. Record **44** (1994)
7. Killeen, P.: Superstition: A Matter of Bias. Not Detectability. Science **199**, 88–90 (1978)
8. Ono, K.: Superstitious behavior in humans. J. Exp. Anal. Behav. **47**, 261–71 (1987)
9. Skinner, B.F.: Superstition in the pigeon. J. Exp. Psychol. **38**(2), 168–72 (1948)
10. Talmont-Kaminski, K.: Malinowski's magic and skinner's superstition: reconciling explanations of magical practices. In: Mental Culture: Classical Social Theory and the Cognitive Science of Religion, pp. 98–109. Acumen Publishing (01 2012)
11. Vyse, S.A.: Believing in Magic: The Psychology of Superstition. Updated Oxford University Press, New York (2014)
12. Vyse, S.: Behavioral variability and rule generation: general, restricted, and superstitious contingency statements. Psychol. Record 487–506 (1991)

Towards Reflective Normative Agents

Nathan Lloyd and Peter R. Lewis

1 Introduction

Artificial Intelligence has been defined as making "computers do the sorts of things that minds can do" [8, p. 1]. Boden's definition evokes questions of whether mind-like faculties can be replicated and to what extent they may modify individual and collective behavior. This interplay between the mind's architecture and behavior observation requires an interdisciplinary approach to developing intelligent agents. This paper discusses the concepts of normative, self-organizing, and reflective agents to propose a framework for capturing the cognitive abilities required to learn, reason, and reflect on norms, not simply follow or comply. The objective is to showcase that reflective normative agents operating in dynamic and complex scenarios can foster the emergence of rich social interactions autonomously, beyond those which do under the assumption that agents are simply maximizing competence or compliance.

Norms can be defined as "the informal rules we live by" [7], ubiquitous within society, and argued as powerful constructs for influencing behavior [24], and also discussed as informal institutional rule types [32, p. 14]. Institutions are systems for organizing and standardizing behavior; their structured rules regulate social behavior and have long been recognized as essential mechanisms for collective action, even when individuals do not share a common purpose [31]. Ostrom's [30] pioneering research in shared common-pool resources discussed sustained institutions and cooperative behaviors through self-organization. Self-organization proposes that individuals create, employ, and modify their institutions to achieve sustainable cooperation [33], mirroring the human potential to change the rules of social interaction [33]. Strides within the social sciences have pushed forward a new constructivist theory of norms [5]. Bicchieri has identified an 'Interdisciplinary Frontier' of norm research [7], compiling theories and empirical science to discuss the significance

N. Lloyd (✉) · P. R. Lewis
Trustworthy AI Lab, Ontario Tech University, Oshawa, Canada
e-mail: Nathan.Lloyd@ontariotechu.ca

C. Elsenbroich and H. Verhagen (eds.), *Advances in Social Simulation*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-031-57785-7_45

of norms within society and discern why individuals may choose to follow social norms. Although this new wave of research provides a map towards the development of normative agents through its formalizations, critical mind-like qualities integral for normative agents are yet to be incorporated.

In discussing the issues of incomplete minds, Lewis and Sarkadi [25] call attention to the failings of many modern AI systems and their inability to reflect upon the social and ethical nature of their decisions. Reflection is a core mental mechanism that motivates the evaluation of beliefs, values, and behavior, essential to assess whether one is congruent with prevailing norms [25] and to reason about the mental states of others. For sustainable self-organization and self-governance, agents require the capacity for reflection [38]. However, it is also essential to understand how norms guide behavior, to formalize how they are represented, learned, activated, and updated [27]. This paper builds upon Bicchieiri's formalized constructivist theory of norms [5], a diverse range of intelligent agent research, and the work of Lewis and Sarkadi [25] to define a reflective normative agent architecture for simulating self-organizing behavior.

2 Norms

First, we introduce Bicchieri's formalization of norms, the conditions to support them, and the necessary properties for norm compliance.

2.1 Components of Norms

In *The Grammar of Society* [5, p. 2], Bicchieri presents a 'constructivist' theory that defines norms in terms of *expectations* and *preferences*. Expectations and preferences are the building blocks of many social constructs; as such, they can be considered integral components for designing and developing artificial social systems. We will first introduce a *personal normative belief*, a concept represented through deontic sentences, and describes what "I believe (I/We) ought to do..." *Social expectations* is an umbrella term encompassing two different types of expectations. The first, *empirical expectations*, is defined as a belief about another's future behavior based on past behavior [6, p. 16–17], written in the form "I expect they're going to..." The second, *normative expectations*, can be described as a second-order belief about another's personal normative belief, commonly expressed through deontic sentences such as "I believe that most people think we ought to do..." An individual's experience informs their expectations, and significant work has looked to explicitly represent them [13]. *Preferences* refer to an individual's disposition to behave in a certain way within a specific context, indicating how expectations alone may not necessarily impact behavior. Preferences may be described as socially unconditional, where others do not influence one's choice, or as conditional, by dependence upon empirical

and normative expectations. It is common to see an additional distinction made in Bicchieri's work that an individual's preferences and expectations are bound to a *reference network*; those who matter to our decision-making processes.

2.1.1 Descriptive Norms

Using Bicchieri's building blocks, a descriptive norm is defined as a pattern of behavior that an individual prefers to engage in on the condition that others within their reference network also engage in it [5]. A descriptive norm describes interdependent behaviors where preferences are conditional upon empirical expectations alone. Following this distinction, descriptive norms drive behaviors such as imitation and coordination, as they are based solely on the behavior of others and not on another's normative expectations.

2.1.2 Social Norms

Whereas descriptive norms are composed of empirical expectations and conditional preferences alone, social norms require the addition of normative expectations. Social norms are interdependent, socially conditional, rely upon social expectations, and require that individuals acknowledge the existence of the normative rules and to which situation they should be applied. Bicchieri [5, p. 11] defines the conditions for a social norm to exist as follows:

> Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. We say that R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq$ such that, for each individual :
>
> *Contingency*: i knows that a rule R exists and applies to situations of type S;
>
> *Conditional Preference*: i prefers to conform to R in situations of type S on the condition that:
>
> *(a) Empirical Expectations*: i believes that a sufficiently large subset of P conforms to R in situations of type S;
>
> and either
>
> *(b) Normative Expectations*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
>
> or
>
> *(b') Normative Expectations with Sanctions*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform, and may sanction behavior.

It is essential to disambiguate the notation of P, which, dependent upon simulation objectives, may represent a reference network or a larger population. This distinction is crucial when applying *behavioral rules*, as one rule may be a social norm in P and not in P'. Bicchieri outlines the conditions for a social norm to exist by requiring a sufficiently large subset of *conditional followers* P_{cf}, a conditional follower, however, merely recognizes the existence of the norm and is said to become a *follower* when

their social expectations are fulfilled. We can then say the norm is *followed* if a sufficiently large subset P_f of P_{cf} meets the conditions of contingency, conditional preference, and social expectations: $P_f \subseteq P_{cf} \subseteq P$.

2.1.3 Not Norms

Finally, in light of often ambiguous and misrepresented terms surrounding the discussion of norms, it is essential to discuss apparently similar concepts. Outlining Bicchieri's building blocks reveals the factor distinguishing normative from non-normative behaviors like customs, habits, shared morals, and religious rules. This factor is interdependency. It is a common pitfall to erroneously group normative and non-normative behaviors together. However, it is essential to note that independent and interdependent behaviors are motivated by entirely different sets of preferences, with independent actions occurring irrespective of what others do.

2.2 Requirements for Norm Competence

Inspired by Bicchieri's definition of norms, recent work by Malle et al. [27] discuss the properties required for an artificial agent with norm competency; here we discuss their propositions in light of Bicchieri's formalizations.

2.2.1 Norm Representation

The language utilized in discussing normative expectations connotes the use of deontic logic. It is common to express these statements as what one should or ought to do or what is obligatory, optional, permissible, or prohibited. Malle et al. [27] propose that normative rules be represented through three distinct categories: prescription, prohibition, and permissions. They suggest a graded ordinal scale to provide granular insight into the demand of the normative rules, which signals the strength of the expectation. Grading normative rules provides an intuitive mechanism for decision-making. The scale provides a weight such that an agent can distinguish to what degree the norms are demanded, i.e., recognizing whether prescriptions are required or suggested. Norms are seldom described as absolutes, their supporting language often possessing a fuzzy and qualitative nature; thus, graded demand becomes appropriate.

2.2.2 Context Sensitivity

Bicchieri's *contingency* condition requires an agent to be aware of a *behavioral rule* and to which *situations* they are applied for their activation. Recognizing a situation implies that an agent must first be able to perceive its environment and then infer

what features within the context activate the norm; situational cues. Situational cues may come from the environment and others within that situation, activating one's beliefs, preferences, and any known accompanying norms. The precise mechanism by which humans perceive contexts and how such contexts trigger the applicable norms is presently unclear [27]; however, this is expected to be computationally demanding for non-humans [36].

2.2.3 Prevalence

A requirement for identifying social norms is to recognize their prevalence within a reference network. Although a definition for social norm followers provides an understanding of the prevalence of a norm, this knowledge cannot be assumed to be available to individual agents. Therefore, prevalence can be calculated or estimated based on what is observed or communicated.

2.2.4 Norm Learning and Updating

Norm learning illustrates the cyclical relationship between external and internal norms [11], where external events influence one's internal representation, and in turn, the internal norm shapes one's behavior. For example, external information may come from explicit instructions, such as signs, verbally communicated rules, or expressing (dis)approval when a norm is either conformed or transgressed. These communications may infer the rule's demand, with stronger sanctions and continued communication of that rule highlighting its significance to the reference network. Observing others' behavior and the consequences of their actions provides another vector from which to learn, but this may be insufficient for learning norms accurately. For example, observing behaviors does not express the individual's desire or motivation; they may be self-interested and act independently of others or be compliant due to pluralistic ignorance. Observations are also limited in realistic situations where agents can only access imperfect information from their surroundings. Thus, to avoid confusing norm-guided behavior and an individual's goals or desires, one can learn from the consequences of actions, whether a behavior is reinforced or sanctioned. However, the enforcement of social norms can vary significantly [5, p. 8] and may be heavily influenced by the interdependency of the reference network [22]. Norm learning thereby describes an observational process to update one's own mental representations and beliefs about others.

3 Towards Reflective Normative Agents

The concepts presented thus far facilitate an agent's ability to acquire, represent, and adhere to norms. However, these concepts primarily ensure competency or com-

pliance, leaving no room for an agent to intentionally violate a norm, which may be advantageous and preferable for achieving an individual's or society's goals [10, 12]. Before Bicchieri's formalizations, Castelfranchi et al. [12] discussed the necessity of *intelligent violations* and the requirement for cognitive agents that may form mental representations of beliefs, goals, and intentions, an aspect missing in recent prior work. Reasoning about these mental representations requires further discussion about different reasoning processes. Bicchieri's work highlights the tendency to focus on deliberation, and higher-level reasoning capabilities like reflection appear absent in the discussion of normative agents.

Open-ended environments entail scenarios where agents may need to learn to coordinate, cooperate, conform, or control one another [3], learning appropriate strategies and self-organizing through their interactions with the environment and one another. Moreover, open-ended situations do away with domain constraints and the specification of narrow problems that may otherwise constrain norms and emergent group behavior. Open-ended, complex environments have seen particular success in developing deep learning agents [29, 35]. Agent-based modeling is a "quintessential tool for open-ended social theorizing" [14], where outcomes (like norms) are socially constructed, emerging organically from social interaction. Indeed, norms are emergent phenomena that come in all shapes and sizes, varying tremendously worldwide due to the complexity of the environment upon which human societies sit. Therefore, it is necessary to model the conditions that initially allow various norms to arise. We propose open-ended situations in which there is no single task. Instead, the emergent norms and behaviors are determined by the initialization of the world and its inhabiting agents.

3.1 Theory of Mind

Social expectations are beliefs about others' behavior and what others believe. To reason about these, an agent must possess models of others that appropriately incorporate their mental states to reason about the prevalence of norms. The cognitive science community has extensively investigated the process of forming mental representations of the goals, beliefs, and preferences of those who are interacted with [16, 41]. The capability to construct mental models of others, known as the Theory of Mind (ToM) [41], is a fundamental aspect of human social intelligence [37]. ToM has multiple orders [19], but social expectations require the second. Zeroth order ToM states that an individual can reason about their knowledge, beliefs, desires, and perceived state of the world but maintain no understanding of the mental state of others [19]. First-order ToM involves recognizing that one and others have desires and beliefs that influence behavior. Second-order ToM recognizes that others may hold beliefs about oneself; essential for normative expectations. The ability to infer the intentions and form beliefs about other agents from observable actions has significant practical applications, particularly for normative agents in cooperative and competitive tasks [28].

Modeling others also provides additional qualities that describe how social norms are adhered to. *Social-image* and *Self-image* are aspects that are congruent with the ToM. These mechanisms facilitate functionalities that help to describe adherence to a social norm in private and public settings [7, 18]. A social image concern refers to an individual's desire to appear in a particular light to others within their reference network and to seek their approval [39]. A social image enables individuals to be aware of how they or others are perceived, determining who may be trustworthy, reliable, and reputable, factors influencing human decision-making [2]. Self-image is discussed as a mechanism to explain why individuals may continue to exhibit normative behaviors in a private setting. Rather than exhibiting idealistic characteristics and behaviors for the benefit of others, individuals choose to adhere to norms to reinforce a positive self-image; to feel good about themselves. Individuals' self and social image concerns infer self-awareness about their behavior and how others perceive them within their reference network. Without explicit coordination protocols, modeling other agents becomes an essential skill for effective collaboration [2] and enables the achievement of common goals with decreased effort [37].

3.2 Diverse Reasoning Capabilities

The social phenomena and properties of norms discussed thus far imply the requirement for a cognitive agent, a type of agent that can emulate the human capacity for memory and problem-solving. This distinction connotes a *stronger* notion of autonomy for agents [42], those characterized by *mentalistic* concepts, able to manipulate and reason upon mental presentations like goals, beliefs, and context. This contrasts the implementation suggested by Malle et al. [27], who define a norm conflict resolution property to obey as many norms as possible. Instead, a cognitive agent would reason about their mental representations to decide whether they should conform or transgress. Cognitive agents, themselves, can be divided into two overarching approaches [25], explicit architectures and emerging cognition from complex systems. The Belief-Desire-Intention architecture is a popular example of explicitly defined cognitive agents [34], which has several extensions to overcome limiting assumptions; a comprehensive review of these architectures and agent decision-making is available in [4]. Systems built using Artificial Neural Networks have also been successful in developing cognitive agents [40]; similar to humans, it is expected that cognition may emerge through model complexity. However, there are concerns that deep learning approaches may develop shortcuts that impede accurate mental representations in ToM research [3].

The standard view for information processing and decision-making is the deliberative route to behavior, a conscious process that weighs each factor against an individual's preferences to determine an outcome. There has been much work in this regard explicitly for normative agents [12, 26], where deliberation is used to consciously reason and decide whether to conform or transgress rather than norm following through some hard-coded filter or goal of maximization. In humans, this

process is costly, requiring time, skill, and effort to systematically weigh all factors and calculate the potential utility of available strategies. As such, the deliberative route to behavior has received criticism for being an over-cited but underused decision-making method [5, p. 4–7]. The over-simplistic view that individuals weigh all their decisions and outcomes is unjustified when considering how some decisions are made instinctively through some reactive or unconscious process. This mode of thinking is dubbed the heuristic route. This information processing method calls upon an in-memory set of rules to prescribe actions based on perceived contextual information, beliefs, desires, and expectations. The heuristic method is strengthened by cognitive shortcuts, where overlapping contextual cues and classes of similar situations allow individuals to generalize or extrapolate one behavioral rule to a new situation. These two modes of information processing are well discussed within the literature, often described as thinking fast and slow [23]; both are thought to be simultaneously occurring in some form or another.

Despite both processes being intuitive for information processing and decision-making procedures, Bicchieri argues that they are incompatible given that the former considers preferences as mental states and the heuristic approach does not [5, p. 6]. Between these two modes lies the dispositional approach, a philosophical tradition that considers beliefs and desires in appropriate circumstances. A dispositional decision-making process infers that individuals will be motivated to act according to their preferences until they are dissatisfied by the outcome of their actions or others, sparking a reflective process. A reflective process is essential for dealing with ambiguity, emergent knowledge, and social context [25]. The dispositional process with reflection reveals how a default behavior (heuristic) may be followed until an individual feels unfulfilled based on their expectations, invoking a conscious and reflective process to adjust rules, beliefs, and goals.

3.3 Reflection

Beyond the requirements for a normative agent, we posit the necessity for a reflective capability within normative agents. Reflection is a higher-level reasoning process than previously discussed, enabling individuals to deliberate on abstract concepts like beliefs, behaviors, and norms concerning actions taken and their outcomes. A reflective agent reasons about their behavior [9], the behavior of others, and the external world. A reflective process is essential for determining whether one's actions were congruent with prevailing norms [25], for dealing with ambiguity, emergent knowledge, and reasoning about social contexts. Lewis and Sarkadi introduce a novel socio-cognitive theory of reflection in artificial intelligence [25], which outlines different tiers of reflective capabilities and the corresponding qualities necessary to attain each tier. Expanding on Hesslow's Simulation Theory of Cognition [20, 21], Lewis and Sarkadi explore the role of simulation and hypothesis testing as reflective processes, which follows seminal cognitive science research that discusses an individuals ability to simulate the behavior of others by adopting

their perspective [15, 17], enabling them to comprehend the intentions or motives of others and respond appropriately in social contexts. In discussing Ostrom's work and the Tragedy of the Commons, Powers et al. [33] call attention to the capability of reflective processes within humans, highlighting how this mechanism enables individuals to "change the rules of the game" and avoid undesirable outcomes; recent work highlights the success reflective self-governance for sustainability [1]. Reflection is a crucial cognitive component for normative agents, essential for reasoning about social expectations, normative rules, models of self, others, the environment, and society, correcting wrong beliefs, and motivating new goals and behaviors.

4 Agent Specification

Situated agents can possess the fundamental capacities of perception, locomotion, and interaction. Normative agents extend this ability to ***observe*** the environment and the actions of those within it, ***form beliefs*** about the behavior of others, ***recognize*** the existence of normative rules and to which ***context*** they are applied, hold a model for ***reference networks***, and ***evaluate*** their adherence to a norm based on achieving their own ***goals***, for example, maintaining satiation. Furthermore, normative agents require the capacity to communicate (directly or through signals) and interpret information from the environment and one another [37]. The open-ended and undefined configurable environments an agent may inhabit make defining an agent's full requirements challenging, with conditional agent requirements dependent on the situation. However, we can consider the agent's and environment's basic properties to motivate emergent behaviors like cooperation, coordination, conformity, and control. Beyond capabilities associated with norms and being situated, a reflective normative agent can ***reflect*** upon its mental representations and update or formulate new normative rules, beliefs, and goals. Reflection can be considered an intentional action triggered during a *shock* following a negative outcome when facing ambiguity in a dispositional reasoning process or an action taken during times of comfort. Given these conditions, we define a normative system as follows. Let $M = (G, E)$ be a formal model where E is the environment, and G is the global population of agents. We can then define the composition of an individual agent $i \in G$ as:

$$(S_i, B_i, D_i, X_i)$$

- S_i is the set of i's observations of their own and other's behavior and state;
- B_i is the set of i's beliefs;
- D_i is the set of i's goals or desires;
- X_i is the set of actions known to i (capabilities);

This notation provides an intuitive and generalizable abstraction for normative agents, with i's knowledge, beliefs, goals, and abilities different from those of another agent j. This formalization maintains a level of abstraction for X_i and D_i, which will remain undefined for open-ended scenarios and open to relevant instantiation for a given situation. It can then be stated that an individual's observations denoted as S_i would inform their beliefs B_i. Given the requirements outlined prior, agent i's beliefs can be stated as follows:

$$B_i = (C_i, P_i, R_i, O_i, E_i, N_i, Q_i, A_i, W_i)$$

- P_i is the set of reference networks known to i, the groups to whom their decisions matter;
- R_i is the set of behavioral rules known by i and to which P they belong;
- O_i is the set of models of agents known to i;
- E_i is the set of empirical expectations i has regarding S_i, R_i, and P_i;
- N_i is the set of normative expectations i has regarding S_i, R_i, and P_i;
- Q_i is the set of personal normative beliefs i has regarding S_i, R_i, and P_i;
- A_i is i's model of self;
- W_i is i's model of the world;

Congruent with Bicchieri's formal model of social norms, an agent's belief model incorporates the aforementioned building blocks for the many social constructs to exist. Beyond these requirements, an agent contains a model of itself A_i, the world W_i, and others O_i to facilitate the cognitive requirements and mechanisms attributed to why individuals may choose to conform or transgress. An agent's model of self may contain characteristics such as risk sensitivity, self-efficacy, or their tendency to seek approval, as well as their self-image and social image. An agent's world model reflects their incomplete knowledge of the state of the world through their perception. Our requirements for a normative agent imply the need to model others, necessitating predictions about another's intentions or goals [5, p. 56]. As such, a model of others, O_i, is the final layer to unpack. Where i's beliefs about another agent j is stated as:

$$O_i^j = (P_i^j, D_i^j, X_i^j, O_i^{O_j}, R_i^j, Q_i^j, A_i^j)$$

- P_i^j is i's beliefs of j's membership to reference networks, the groups to whom i believes impacts j's decisions;
- D_i^j is i's beliefs of j's goals or desires;
- X_i^j is i's beliefs of actions known to j's (capabilities);
- $O_i^{O_j}$ is i's beliefs about j's model of others;
- R_i^j is the set of behavioral rules i believes j to know;

- Q_i^j is i's beliefs of j's personal normative beliefs;
- A_i^j is i's beliefs of how j perceives i;

This specification does not explicitly capture the processes for learning, updating, and reflection; these will be explored in future work via operationalization.

5 Conclusion

Our exploration into various aspects of reflective normative agents has shed light on the complexity and importance of understanding human behavior and decision-making processes. Through extending Bicchieri's formalizations, we have delved into the essential components that a normative agent can possess. These include incorporating mental representations of self, others, environment, and society, considering multiple modes of reasoning, and the significant role of reflection in shaping an agent's actions. Furthermore, we have underscored the importance of creating dynamic and complex environments that are open-ended, allowing for the emergence of rich self-organizing behavior. To appreciate the complexity of reflective normative agents, it will be necessary to situate agents in conditions that motivate social constructionism. This highlights the need for a unified testbed containing sufficiently complex and open-ended scenarios to study agent capabilities. Through our exploration, we have come to appreciate the multifaceted nature of normative and cognitive agents, recognizing the complexity of human behavior and the need for nuanced modeling. By addressing these requirements and leveraging the power of reflection, we can develop more advanced agents that demonstrate autonomous decision-making and exhibit the richness of social constructionism.

References

1. Aishwaryaprajna, Lewis, P.R.: Exploring intervention in co-evolving deliberative neuro-evolution with reflective governance for the sustainable foraging problem. In: Artificial Life Conference Proceedings, p. 140 (2023)
2. Albrecht, S.V., Stone, P.: Autonomous agents modelling other agents: a comprehensive survey and open problems. Artif. Intell. **258**, 66–95 (2018)
3. Aru, J., Labash, A., Corcoll, O., Vicente, R.: Mind the gap: challenges of deep learning approaches to theory of mind. Artif. Intell, Rev (2023)
4. Balke, T., Gilbert, N.: How do agents make decisions? a survey. J. Artif. Soc. Soc. Simul. **17**(4), 13 (2014)
5. Bicchieri, C.: The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press (2005)
6. Bicchieri, C.: Norms in the Wild: How to Diagnose, Measure, and Change Social Norms. Oxford University Press (2017)
7. Bicchieri, C., Dimant, E., Gelfand, M., Sonderegger, S.: Social norms and behavior change: the interdisciplinary research frontier. J. Econ. Behav. Organ. **205** (2023)
8. Boden, M.A.: AI: Its Nature and Future. Oxford University Press (2016)

9. Brazier, F., Wijngaards, N.: Designing self-modifying agents. In: Gero, J. (ed.) Computational and Cognitive Models of Creative Design V, pp. 93–112. University of Sydney, Key Centre of Design Computing and Cognition (2001)
10. Burth Kurka, D., Pitt, J., Lewis, P.R., Patelli, A., Ekárt, A.: Disobedience as a mechanism of change. In: 2018 IEEE 12th International Conference on Self-Adaptive and Self-Organizing Systems (SASO), pp. 1–10 (2018)
11. Castelfranchi, C.: A cognitive framing for norm change. In: Dignum, V., Noriega, P., Sensoy, M., Sichman, J.S. (eds.) Coordination, Organizations, Institutions, and Norms in Agent Systems XI, pp. 22–41. Springer International Publishing, Cham (2016)
12. Castelfranchi, C., Dignum, F., Jonker, C.M., Treur, J.: Deliberative normative agents: principles and architecture. In: Goos, G., Hartmanis, J., van Leeuwen, J., Jennings, N.R., Lespérance, Y. (eds.) Intelligent Agents VI. Agent Theories, Architectures, and Languages, vol. 1757, pp. 364–378. Springer, Berlin (2000)
13. Cranefield, S.: Agents and expectations. In: Coordination, Organizations, Institutions, and Norms in Agent Systems IX: COIN 2013 International Workshops, COIN@ AAMAS, St. Paul, MN, USA, May 6, 2013, COIN@ PRIMA, Dunedin, New Zealand, December 3, 2013, Revised Selected Papers 16, pp. 234–255. Springer (2014)
14. Devereaux, A., Wagner, R.E.: Agent-Based Modeling as Quintessential Tool for Open-Ended Social Theorizing. Tech. Rep. 19-07, George Mason University, Fairfax (2019)
15. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mind-reading. Trends Cogn. Sci. **2**(12), 493–501 (1998)
16. Gopnik, A., Wellman, H.M.: Why the child's theory of mind really is a theory. Mind Lang. **7**(1–2), 145–171 (1992)
17. Gordon, R.M.: Folk psychology as simulation. Mind Lang. **1**(2), 158–171 (1986)
18. Gross, J., Vostroknutov, A.: Why do people follow social norms? Curr. Opin. Psychol. **44**, 1–6 (2022)
19. Hedden, T., Zhang, J.: What do you think i think you think?: Strategic reasoning in matrix games. Cognition **85**(1), 1–36 (2002)
20. Hesslow, G.: Conscious thought as simulation of behaviour and perception. Trends Cogn. Sci. **6**(6), 242–247 (2002)
21. Hesslow, G.: The current status of the simulation theory of cognition. Brain Res. **1428**, 71–79 (2012)
22. Horne, C.: Explaining norm enforcement. Rational. Soc. **19**(2), 139–170 (2007)
23. Kahneman, D.: Thinking. Fast and Slow. Farrar, Straus and Giroux, New York (2011)
24. Kalkstein, D.A., Hook, C.J., Hard, B.M., Walton, G.M.: Social norms govern what behaviors come to mind-And what do not. J. Person. Soc, Psychol (2022)
25. Lewis, P.R., Sarkadi, S.: Reflective artificial intelligence. Mind. Mach. **34**(2),1–30. Springer (2024). https://doi.org/10.1007/s11023-024-09664-2
26. López y López, F., Luck, M., d'Inverno, M.: A normative framework for agent-based systems. Comput. Math. Organ. Theory **12**(2-3), 227–250 (2006)
27. Malle, B.F., Bello, P., Scheutz, M.: Requirements for an artificial agent with norm competence. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 21–27. AIES '19, Association for Computing Machinery, New York (2019)
28. Matiisen, T., Labash, A., Majoral, D., Aru, J., Vicente, R.: Do deep reinforcement learning agents model intentions? Stats **6**(1), 50–66 (2023)
29. Open Ended Learning Team, Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., McAleese, N., Bradley-Schmieg, N., Wong, N., Porcel, N., Raileanu, R., Hughes-Fitt, S., Dalibard, V., Czarnecki, W.M.: Open-ended learning leads to generally capable agents (2021)
30. Ostrom, E.: Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press (1990)
31. Pitt, J., Diaconescu, A., Bollier, D.: Technology for collective action [special section introduction]. IEEE Technol. Soc. Mag. **33**(3), 32–34 (2014)
32. Polski, M.M., Ostrom, E.: An Institutional Framework for Policy Analysis and Design (1999)

33. Powers, S.T., Ekárt, A., Lewis, P.R.: Modelling enduring institutions: the complementarity of evolutionary and agent-based approaches. Cogn. Syst. Res. **52**, 67–81 (2018)
34. Rao, A.S., Georgeff, M.P.: Bdi agents: from theory to practice. In: International Conference on Multiagent Systems (1995)
35. Samvelyan, M., Kirk, R., Kurin, V., Parker-Holder, J., Jiang, M., Hambro, E., Petroni, F., Küttler, H., Grefenstette, E., Rocktäschel, T.: Minihack the Planet: A Sandbox for Open-ended Reinforcement Learning Research (2021)
36. Scheutz, M., Malle, B.: Moral Robots, pp. 363–377. Routledge/Taylor & Francis Group (2017)
37. Sclar, M., Neubig, G., Bisk, Y.: Symmetric machine theory of mind. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 19450–19466. PMLR (2022)
38. Scott, M., Pitt, J.: Interdependent Self-Organizing Mechanisms for Cooperative Survival. Artificial Life, pp. 1–37 (2023)
39. te Velde, V.L.: Heterogeneous norms: Social image and social pressure when people disagree. J. Econ. Behav. Organ. **194**, 319–340 (2022)
40. Volzhenin, K., Changeux, J.P., Dumas, G.: Multilevel development of cognitive abilities in an artificial neural network. Proc. Natl Acad. Sci. **119**(39), e2201304119 (2022)
41. Woodruff, G., Premack, D.: Intentional communication in the chimpanzee: the development of deception. Cognition **7**(4), 333–362 (1979)
42. Wooldridge, M., Jennings, N.R.: Agent theories, architectures, and languages: a survey. In: Wooldridge, M.J., Jennings, N.R. (eds.) Intelligent Agents, pp. 1–39. Springer, Berlin (1995)

The Curse of Possibilities

Nicolas Payette

Agent-based models have inputs and outputs:

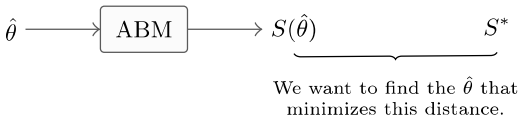
Some models are meant to represent the world. In those cases, we would like the inputs of our model to correspond to some sort of initial condition in the real world and the outputs to correspond to an outcome that can be measured in both world and model. In practice, inputs and outputs are often (but not always) sets of numeric values: a set of parameters θ for the inputs and a set of summary statistics S for the outputs.

The values of some parameters might be known to us, whether it's the hold size of a particular fishing vessel, the maximum running speed of a grey wolf or the base interest rate of the Bank of England in May 2023. We can change those values if we want to explore counterfactual scenarios, but they are still bound by empirical data or by our assumptions. Other parameters, however, are "free": we do not have any direct way of establishing their values. Sometimes that's because the parameter's empirical value is difficult to measure (e.g., the "attraction rate" of a fish-aggregating device) or because the parameter concerns a more theoretical part of our model, for which it is difficult to pluck a value directly from the real world (e.g., the learning rate of a reinforcement-learning agent).

Even if we can't estimate free parameters directly, all is not lost: we can try using what we know about outcomes in the real world to work our way back to a set of parameter values. This task, usually called *calibration* in the ABM community, can be tackled in different ways [10], but the general idea is to find a vector of estimated parameters $\hat{\theta}$ that minimizes the distance between model output for those parameters $S(\hat{\theta})$ and observed summary statistics S^* (i.e., the error of the model):

N. Payette (✉)
School of Geography and the Environment, University of Oxford, Oxford, UK
e-mail: nicolas.payette@ouce.ox.ac.uk
URL: https://www.geog.ox.ac.uk/

C. Elsenbroich and H. Verhagen (eds.), *Advances in Social Simulation*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-031-57785-7_46

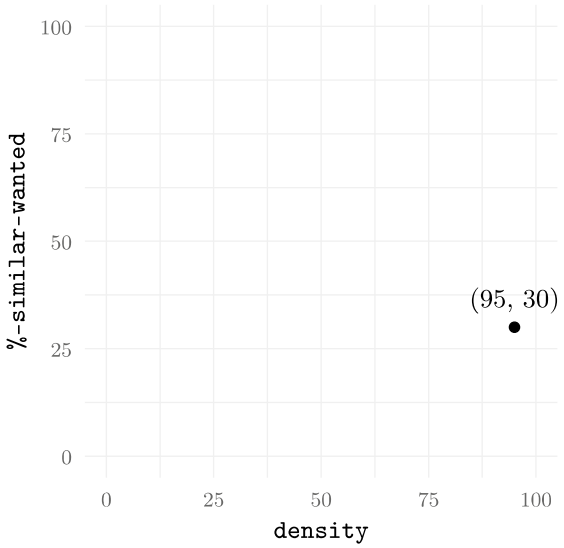


Calibration is a tricky topic, and the best way to do it varies from model to model [1], but this simplified picture gives us a way to think about the various mathematical spaces that we are interacting with during the modelling process. We will develop these ideas by looking at the NetLogo [14] version [12] of the Schelling model of segregation [8, 9]. *Caveat emptor*: the original Schelling model is a conceptual model, not meant to be empirically calibrated.[1] It is nevertheless useful as an example because it is both simple and well known.

NetLogo's model library contains a Schelling-like model using a movement rule that's different from the original.[2] It is called *Segregation* and has two parameters, listed here with their default values:

Parameter	Value	Description
Density	95	The percentage of grid cells occupied by agents
%-Similar-wanted	30	The agents' desired percentage of neighbours of the same colour as them

That gives us our initial θ, which in this case is simply a tuple of values: (95, 30). It's easy to see how each possible θ gives us a point in the *parameter space* of the model:



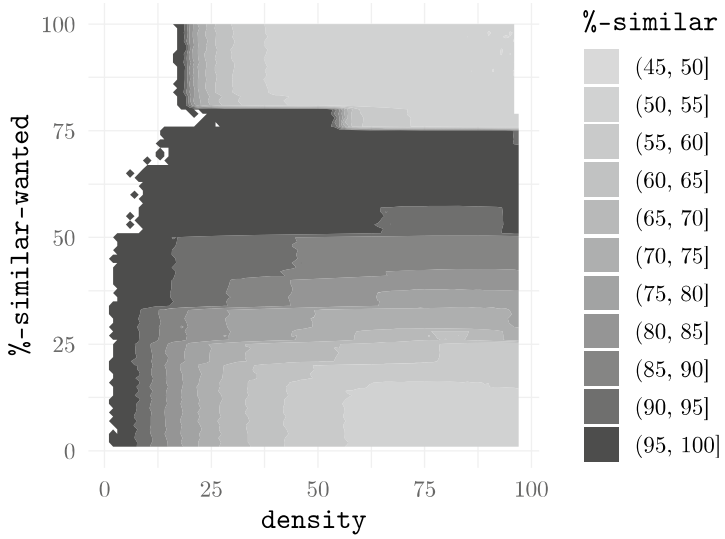
Parameter space is not the only space we're dealing with: there is also the *state space* of the model. In the case of the *Segregation* model, each agent has three

[1] There are plenty of Schelling-inspired empirical models out there. See [15] for a recent example.

[2] A detail which is not completely unrelated to the point that we will be trying to make in this paper.

relevant[3] variables: `xcor`, `ycor` and `color`. Each of these variables, for each agent, adds a dimension to the model's state space.[4] Running a simulation can be thought of as following a trajectory through the model's state space.

Summary statistics (S) are just a pragmatic way of looking at the model's state space, making it easier to interpret and relate to the real world (since we probably don't have access to the whole state space of our empirical target anyway). In the *Segregation* model, the one relevant summary statistic is `%-similar`. It's defined as the sum of the count of similar neighbours for each agent, divided by the sum of the count of all neighbours for each agent. Imagine that we have an empirical S^* (even though we don't). Since our parameter space is so small, we can brute force our way through it and find the $\hat{\theta}$ that will produce S^*. Easy, right? Look at the results:



Two things should be noticed from the above plot. First, different regions of the parameter space can produce very similar results. If you're looking for a `%-similar` between 70 and 75, many combinations of `%-similar-wanted` and `density` will give you that, so can't know what particular combination occurs in the real world. That's the problem of parameter *identification*. If this was a proper empirical model, we could avoid it by obtaining `density` from data. But then again, if this was a proper empirical model, we would have more than two parameters. Parameter identification is a long-acknowledged challenge [4, 7] that we are not going to confront head-on but that should be kept in the back of the reader's mind while we move along.

The other—perhaps even more obvious—thing is that we are missing results for very low values of `density`, especially when `%-similar-wanted` is high.

[3] NetLogo defines various other agent variables that complicate things a bit, like `heading`, `label`, or even `pen-size`, but those can be safely glossed over for now.

[4] Things get hairier when a model dynamically creates or remove agents, but let us not think about that too hard.

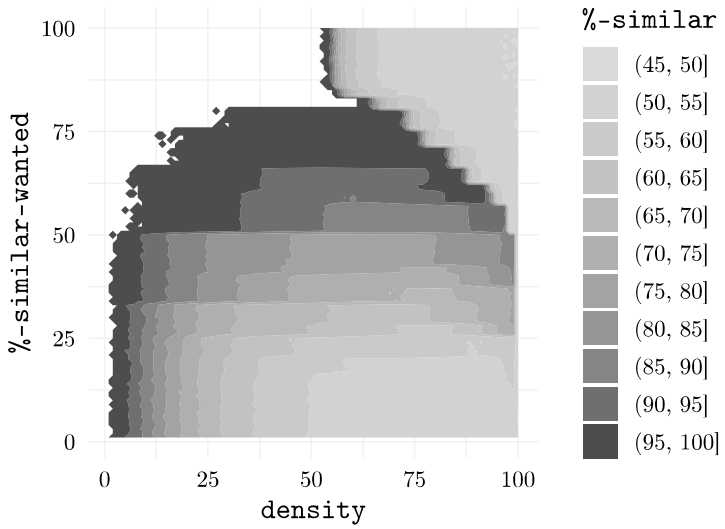
That is a consequence of the way `%-similar` is defined in the model: it divides the number of similar neighbours by the total number of neighbours. At low densities, it's possible for every agent to be completely isolated, in which case there are no neighbours at all and we end up with a division-by-zero error.[5] This explains why the `density` slider has a lower bound of 50% in the model's interface. Not every region of the parameter space is relevant to a model's analysis, but we tend to shy away from pushing against those boundaries because of the complications they introduce.

Speaking of complications: we have pretended so far, for the sake of simplicity, that *Segregation* is a two-parameter model, but is it really? One nice thing about NetLogo is that it comes with a predefined world ready to be inhabited by your agents (or "turtles", in its parlance): a continuous two-dimensional space divided in "patches". You can jump right into modelling without getting bogged down in the details because some choices have already been made for you. The world, by default, measures 33×33 patches and the space is toroidal (i.e., doughnut shaped; you can wrap around vertical and horizontal edges). In the *Segregation* model, world size has been set instead to 51×51 and world wrapping kept on. Does it matter? The results might not be too different on a 33×33 grid, or even 100×100, but what about 3×3, or even 50×1? Topology might matter too: in a non-wrapping world, corner agents would only have three potential neighbours instead of eight. Trivial details, maybe—or fun things to try, depending on how you see it—but both world-size and topology are very much a part of the model's parameter space even if you choose not to explore it. We tend to think of model parameters as either numeric (integer or floating-point values), logical (boolean values) or sometimes enumerated (i.e., one of a set of possible choices, often implemented as strings), but this is just the tip of the iceberg.

We mentioned above that NetLogo's *Segregation* model uses its own movement rule: an unhappy turtle picks a random heading and moves forward until it finds a free spot (whether or not it's happy there). Schelling's original rule is that "an individual discontent with his own neighborhood moves to the nearest vacant spot that surrounds him with a neighborhood that meets his demands" [9]. Running the model with Schelling's rule gives us results that are similar but not quite the same:[6]

[5] I had to modify the *Segregation* model to output "`NA`" instead of crashing when running it through *BehaviorSpace*. I ran 50 iterations per condition and report the mean, excluding runs that resulted in NA values.

[6] This is not the point but I can't help to notice how moving to the closest spot that satisfies an agent's demands (which can very well be a spot with no one else around) is much more likely to generate NA values than moving the the first free spot in a random direction (which is much more likely to be right next to someone else).



This is not surprising. Both movement rules support Schelling original insight about how relatively low homophily can still lead to high segregation, so maybe it doesn't matter all that much. But, again, put yourself in the shoes of an empirical modeller trying to calibrate their model: which rule should you use? Perhaps you should try both: make the rule an extra parameter of your model. Congratulations: it now takes twice as long to explore your model's parameter space and your identification problem just got worse. If both rules can generate S^*, which one should you pick? What about the very large number of other rules that could have been used instead? And that's just movement. We haven't said anything on model initialization,[7] updating order, or even how "distance" is defined.[8] All those things are also parameters.

Note how the distinction between "fixed" and "free" parameters applies here too. Suppose you tell me that you don't need a parameter for a certain agent behaviour because we know with great confidence how real-world agents behave in such situations. That's great, but that's no different from knowing the population density because you are modelling a real-world city: it's still a parameter, it's just that you happen to know the value for that particular one.

The point we have been building towards is this: the distinction between parameter space and program space (or "model space", "design space", "specification space") is arbitrary. It is a scary thing to contemplate: it means the parameter space of any model can never be fully explored. We are cursed with effectively infinite possibilities and in order to achieve anything we are forced to make choices, often based on the flimsiest of reasons, which carve down model space to a tiny shell of its expansive self and banishes the rest of it to the realm of the unthinkable.

---

[7] To quote Schelling himself: "It might make sense to distribute the blank spaces evenly, but I let them be determined at random. (It makes some difference.)" [9, 155].

[8] I realised after the fact that Schelling uses Manhattan distance whereas I implemented his rule using Euclidean distance. I do not know if it makes any difference.

Let's take a breath and think about the consequences. Nothing has been said in this paper that is not already known and agent-based modelling is alive and well, so maybe we just keep on keeping on? That's not an unreasonable suggestion. The curse of possibilities is a hump we can get over, and I believe that is something we see in our personal trajectories as modellers. As a beginner, you are unaware of the problem. You just want your model to work, and once it does, you're quite happy to just explore the parameter space as it is. As you gain experience, you start seeing all the possibilities, agonizing over every choice. It's not unlike the old fable of the fox and the cat: the fox is bragging to the cat about how many cunning escape tricks he knows, whereas the cat has only one trick which is to climb up a tree. When the hounds come, the cat immediately gets up to safety but the fox fails to decide which of his many tricks to use and gets caught. With experience, we learn to go back to slightly more cat-like behaviour. We develop intuitions about what works and what doesn't and we become acutely aware of how limited resources (both human and computational) are. Deadlines are looming and papers need to get out so we just call the shot: this is my model, those are its parameters; I'm going to calibrate it and do my sensitivity analysis and everything will be fine. What else could I possibly do?

If there is one call to action in this essay, it is this: at least be explicit. Do not turn a blind eye and hope others will too. It's OK to make arbitrary choices; there is no way around it, but those choices still matter and need to be documented. Edmund Chattoe-Brown [2] conducted a review of papers published in JASSS between 1998–2022 using the terms "random movement" or "move randomly". He found that the majority of those papers did not fully explain what that meant and that those that did all meant different things. Yet, as he shows using NetLogo's *Wolf-Sheep Predation* model [13], and as we have seen with Schelling, movement rules affect results.

So should we stop here, at the acceptance stage of grief, and be content in acknowledging the arbitrariness of our modelling choices? I believe we can do better and that the way forward is not just to be explicit, but to be *formally* explicit.

Let's go back to Schelling and consider movement rules. Once we have decided that agents should move to a free spot if they are unhappy, we can think of any possible movement rule as a *function*: it takes the current state of the model as an input (i.e., the position and colour of every agent) and outputs the coordinates at which the agent wishes to move.[9] Since possible model states are finite and possible choices also finite, there is a finite (though massive) number of possible mappings between them. And that goes for most behaviours in an agent-based model: if we formally define a behaviour as a function with explicitly typed inputs and outputs, we can get the computer to generate possible implementations of those functions. This becomes part of the calibration process: we can automate the exploration of functional space, just like we automate the exploration of numeric parameter dimensions.

This possibility has recently received attention in the agent-based modelling community under the "Inverse Generative Social Science" (iGSS) moniker [3]. That

[9] NetLogo subtly encourages modellers to change the state of their agents directly—the *Segregation* movement rule is a good example of that. It's arguably a very intuitive way of thinking, but it also deprives us of the abstraction opportunities that come from modelling behaviours as pure functions.

methodology has even been applied to the Schelling model itself, in a much more sophisticated way than what I am discussing here [5]. The idea that we can use a computer program to generate computer programs is not new. It was first really developed by John Holland in the 1970s [6] but, like many things in computer science, it dates back to Turing [11].

So if it's an old idea and ABM practitioners are already using it, what am I advocating for? I think that we are on the right track, but there are two things we still need in order to make more progress. The first and most pressing thing is a proper conceptual framework. We can talk about collapsing spaces and how it's all really just the same, but that only muddies the waters. A proper mathematical treatment of these questions would set the stage for the second thing which I believe is needed: better technical tools. At this point you might be recoiling in horror and asking: you are not really suggesting yet another ABM platform, are you? A fair question. I am agnostic as to whether a new platform is required, but there are a few features that I think we are missing.

We need to be able to formally identify the parts of our model that we consider to be parameters and the parts that we don't. This might seem counter-intuitive, as I have just advocated that this delineation is arbitrary, but *au contraire*: it just makes it all the more important to be clear about it. These parameters could be defined as scalar types (i.e., numeric, logical or enumerated) or as functions with their input and output types. Furthermore, for each of those parameters, it should be explicitly stated whether they are fixed or free, and in the latter case what their bounds are. The same care should be taken in defining summary statistics for the model, with the possibility of stating which ones have a corresponding empirical target and which ones are just informational. Again, nothing new here: this is all just good practice. It's also something that our current tools are not very good at helping us with: it requires discipline and experience when, ideally, it should happen almost effortlessly.

One way to think about all this is that we should provide metadata about our model. It would help with documentation, but it would also help with *automation*. Once we know all the parameters with their bounds and all the summary statistics with their targets, we can automate calibration and sensitivity analysis. Once we know the functional definition of our model's behaviours, we can automate the generative search for possible implementations. Extensibility and interoperability should both play an important role in this brave new world. We cannot presume that current methods for calibration, sensitivity analysis and, especially, function generation are the be all and end all and could all be packaged in a static library. New plugins should be easy to write and external libraries (especially those written in Julia, Python and R) easy to leverage.

With such a tool in hand, maybe we could be a bit more like the fox again.[10] We could afford to embrace a much wider space of possibilities without getting paralysed by it. We could focus on the things we really know how to model (there will always

[10] I seriously considered *From cat to fox to cat, then back to fox again* as an alternate title for this essay.

be at least a few) and let the machine deal with the things we don't, in a more efficient way that we ever could. It would not solve all our problems, but it would partly lift the curse of possibilities.

References

1. Carrella, E.: No free lunch when estimating simulation parameters. J. Artif. Soc. Soc. Simul. **24**(2), 7 (2021)
2. Chattoe-Brown, E.: All the right moves? Systematically exploring the effects of random travel in agent-based models. In: Advances in Social Simulation. Proceedings of the 17th Social Simulation Conference, 12–16 Sept 2022. Springer Proceedings in Complexity (2023)
3. Epstein, J.M.: Inverse generative social science: backward to the future. J. Artif. Soc. Soc. Simul. **26**(2), 9 (2023)
4. Fisher, F.M.: The Identification Problem in Econometrics. McGraw-Hill (1966)
5. Gunaratne, C., Hatna, E., Epstein, J.M., Garibay, I.: Generating mixed patterns of residential segregation: an evolutionary approach. J. Artif. Soc. Soc. Simul. **26**(2), 7 (2023)
6. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press (1975)
7. Manski, C.F.: Identification Problems in the Social Sciences. Harvard University Press (1995)
8. Schelling, T.C.: Models of segregation. Am. Econ. Rev. 488–493 (1969)
9. Schelling, T.C.: Dynamic models of segregation. J. Math. Soc. **1**(2), 143–186 (1971)
10. Thiele, J.C., Kurth, W., Grimm, V.: Facilitating parameter estimation and sensitivity analysis of agent-based models: a cookbook using NetLogo and R. J. Artif. Soc. Soc. Simul. **17**(3), 11 (2014)
11. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236), 433–460 (1950)
12. Wilensky, U.: NetLogo Segregation Model. Northwestern University, Center for Connected Learning and Computer-Based Modeling (1997)
13. Wilensky, U.: Wolf Sheep Predation Model. Northwestern University, Center for Connected Learning and Computer-Based Modeling (1997)
14. Wilensky, U.: NetLogo. Center for Connected Learning (1999)
15. Zuccotti, C.V., Lorenz, J., Paolillo, R., Rodríguez Sánchez, A., Serka, S.: Exploring the dynamics of neighbourhood ethnic segregation with agent-based modelling: an empirical application to Bradford. UK. J. Ethnic Migrat. Stud. **49**(2), 554–575 (2023). https://doi.org/10.1080/1369183X.2022.2100554

Using Survey Data to Develop Agent-Based Models of Spatial Segregation

Daniel Schubert

1 Introduction

With the increased immigration of refugees to Germany since 2015, there has been a strong increase in the proportion of people with a migration background as well as a strong diversification of immigrants' references to origin. Immigration and diversification are changing the ethnic composition of the population, especially in cities, and thus driving processes of ethnic segregation. The phenomenon of segregation is a well-known topic but in its theoretical foundations very limited. In this short paper, I will point out a theoretical approach for an agent-based simulation of segregation pattern and investigate how different levels of tolerance influence the outcome of the model. To develop the different level of tolerance, data of the German General Survey (ALLBUS) 2006 will be analysed and implemented in an agent-based model.

2 Theoretical Background

The theoretical foundation of segregation is usually seen in the model of Schelling. Schelling ([1]: 139) writes that the choice of neighbourhood is a choice of neighbours. This means for people who want a specific social group in their neighbourhood force to move in these areas because the majority is like a hint for specific characteristics. An explanation he uses is that other people have seen skin colour as a signal and the people now reflect the signal. Schelling ([1]: 141) acknowledges that there are always segregated neighbourhoods, but the characteristics can differ. In the US usually ethnic segregation is researched by skin colour and that it is difficult to find

D. Schubert (✉)
Sociology/Urban and Regional Sociology, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany
e-mail: daniel.schubert-p3r@rub.de

C. Elsenbroich and H. Verhagen (eds.), *Advances in Social Simulation*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-031-57785-7_47

neighbourhoods that are neither 75% populated with people of white or black skin colour. In a comparison, Schelling writes that it is also difficult to find areas that have a balance that lasts long enough. Most segregation studies focus on ethnic segregation but there are other forms of segregation like social or demographic. In this short paper I will focus on ethnic segregation.

Schelling ([1]: 141) admits that his model does not explain how segregation occurred. In this context, Schelling assumes that an analysis of the majority situation can only happen locally and that each group strives for this numerical superiority. When this is achieved, they try to segregate themselves from the smaller group. To understand segregation, the researcher needs to analyse the incentives that motivate or sustain behaviour. However, no equilibrium is sought by both groups, since the disappearance of a minority leads to complete segregation. At the same time, complete segregation is a stable condition, while in other conditions a shift in the mix is always possible ([1]: 142). The question that arises is how these mixtures are influenced by individual decisions. According to Schelling ([1]: 144), some rules are used to obtain legitimacy for certain behaviours.

Schelling ([1]: 138) acknowledges that he leaves out two processes in his model. One is the organised actions, whether these are legal/illegal, forced or merely exclusionary, or subtle or blatant. The other is the effect of socio-economic aspects. By this, Schelling means above all processes that separate the rich from the poor or the better educated from the less educated ([1]: 139). In our contemporary understanding, we would call this process social segregation. He acknowledges that race correlates closely with income and thus influences residential choice. That means that the residential segregation is a result of these residential choices and can be seen as manifestation of social distance. He assumes that even if race is ignored, segregation occurs.

The question is how segregation arises and what theoretical foundation can be used for Schelling's model? This aspect is not explained in any detail in numerous models; instead, Schelling's simple description is used as a theoretical foundation (e.g. [2, 3]). In the following, I would therefore like to discuss the established-outsider configuration according to Elias and Scotson [4].

The negotiation of norms is carried out through the process of communicating the action taken, i.e. the articulated advocacy. However, the exercise of the new norm leads to a demarcation between the behaviour of the opponents of the new norm and reality. Thus, a conscious demarcation is focused on. Elias and Scotson [4] refer to such a figuration as an established-outsider configuration.

Established people, according to Elias and Scotson ([4]: 9), are characterised by the conviction of a group charism in which all members of the group participate. With the group charism comes a strong "we" ideal and the belief that one's self belongs to a group of people who have a higher value (Elias and Scotson 1993: 9). In this context, the self-image is shaped by the most positive subgroup of the establishment group. The counterpart to this is group shame. Group shame is characterised by a group being ascribed the "worst" characteristics of its "worst" subgroup ([4]: 13). This results in a social devaluation, of the outsiders ([4]: 9). The stigmatisation of groups is usually linked to collective ideas about certain groups. Through the

ideas, the behaviour of the stigmatising group becomes excusable, since it is not the stigmatising group that has endowed the stigmatised group with the characteristic, but this characteristic comes from higher powers. Thus, the characteristic is used as a sign of inferiority or badness. This symbol not only has the function of relieving the stigmatising group of the burden of guilt, but also of defending the existing balance of power ([4]: 32p.). Segregation is a mixture of such structures. On the one hand side person would avoid moving into districts where they are in the position of the minority and on the other hand side leave the district if the neighbourhood changes. Segregation thus arises primarily out of the need to distinguish oneself from others and to maintain a positive group perception. The theories of Elias and Scotson [4] can be used as a theoretical foundation of the processes described by the segregation model of Schelling. The problem Schelling [1] mentioned that this model is missing a theoretical basis can be solved when using this theory. Even if individuals do not want to separate themselves, the desire for at least 50% of the same social group in the environment leads to the emergence of segregation patterns.

In research on segregation, ABM is also used regularly. In an international comparison, many studies on segregation have already been conducted (e.g. [5]). Therefore, only a few selected studies will be discussed here. Hatna and Beneson [3] were able to use ABM to generate a real-world picture of the distribution of religious communities in two Israeli cities by using census data from 1995 and taking into account residential preferences in relation to one's own religion and a similar prestige of the residential environment. Their models were able to replicate real-world mixed neighbourhoods. Another example is Liu et al. [6]. Liu et al. [6] considered actors who could influence neighbours in their ABM and showed that this changed relocation behaviour. Zuccotti et al. [7] showed that segregation is influenced by socio-economic status and ethnicity. These selected models show that ABMs can be used to replicate real-world segregation patterns. At the moment there are no segregation models developed for the German context. Moreover, there is another aspect that is neglected. Working with real-world data and ABM on segregation is therefore not a new approach. The key difference is that this model uses data on behaviour and is not intended to replicate real-world patterns, but to check whether segregation is occurring and whether a stable model is developing. Building on the theory of Elias and Schelling, this model will look at how the model changes when real world data is used for the model. Here, the tolerance values function as an indicator of whether people accept other group members in their neighbourhood.

3 Hypothesis and Research Question

Agent-based modelling (ABM) are simulation studies that provide a better understanding of the course of segregation processes. Through modelling, it is possible to observe how individual characteristics generate different macro phenomena ([8]: 453). One way to make the spatial segregation model more realistic is to use survey data. To simplify the model, it is assumed in the simulation study that the agents

perceive their neighbours and adapt their actions accordingly. What preferences exist with regard to neighbourhoods and how do these affect segregation patterns? This is the research question that is the focus of this presentation.

With this theoretical foundation of Schelling [1] and Elias and Scotson [4] the hypothesis arises that persons reject neighbourhoods in which they are in the minority position and would avoid them. This a However, this makes the behavioural assumption of the actors more complex. It is therefore no longer only decisive that a certain number of the same group are in the neighbourhood, but also how many members of the foreign group there are. This behavioural assumption is more complex and the question that arises from this is how this more complex assumption affects the segregation process. This assumption of behaviour is supported by the model assumptions of Schelling and results from the theory of Elias and Scotson (1993). The specific tolerance values should arise from empirical data. The research question is how these values effect the model outcome.

4 Methods

To investigate this hypothesis, I proceeded in two steps. In the first step, the data on the preference over the ethnic composition of the neighbourhood from the ALLBUS 2006 were analysed. There, in accordance with Farley et al. [9], the consent to the different composition of neighbourhoods was asked. The first step is a descriptive evaluation of this vignette-like designed questions. In the second step, a model was programmed in Netlogo [10]. The model replaces this assumption with the more realistic preference distribution determined empirically from the ALLBUS data.

5 How the Model Works

The basic structure does not differ much from the segregation model of Wilensky [11], only the specifications made make it significantly more complex than the original model. These changes will be discussed in the following. First, the model assumes two distinguishable groups. For convenience, the agents have been coloured blue and red. This colouring can be perceived by the individual agents. The two groups are randomly distributed, just as is the case with Schelling. The neighbourhood in this case is therefore based on pure perception by the agents. The social identity is based on one's belonging to a social group. In the following, we will discuss which settings can be made in the model.

It is possible to vary the density of the agents. The density of the agents can be varied between 0 and 100% via a controller. Density is understood as the number of free fields on the model area and where agents can move to. In this way, the assumptions of Schelling can be used. In addition, there are two controllers for the

neighbourhood preferences. In the model the share of similar and unsimilar neighbours can be changed through sliders. Therefore, in this text the term sliders is used. The %-similar-wanted slider specifies how many agents of the same social group are desired in the neighbourhood. A second slider (%-unsimilar-wanted) specifies how many different neighbours are desired in the neighbourhood. Both sliders can be set to a value between 0 and 100.

Slider Noise adds a disturbance term to the model that specifies a random value and thus makes the model more realistic. In addition, monitors can be used to observe the concrete number of members in each group, how many agents are happy and how similar the agent groups are. To begin with, there are the variables of how high the average value of equal neighbours desired by all agents is and what percentage of all agents are unhappy.

Table 1 shows the properties of the agents at the beginning.

The values of the belonging to one of the social groups, the tolerance-level, similar-wanted and unsimilar-wanted are drawn randomly.

Table 1 Properties of the agents at the beginning

Property	Value	Concept
Happy?	True/false	If happy is false the agent moves and look for a better place nearby
Similar-nearby	Numeric	How many neighbouring patches have a household with my group?
Other-nearby	Numeric	How many have a household of another social group?
Total-nearby	Numeric	Sum of other and similar variables
Similar-nearby-fraction	Numeric	Fraction of neighbours who have the same social group than me
Attractiveness-now	Numeric	To compute and store current attractiveness once moving is considered
Patch-now		To remember current patch before household starts moving
Similar-nearby-opt		Optional new patch: how many similar are nearby
Other-nearby-opt		Optional new patch: how many household of another social group are nearby
Total-nearby-opt		Optional new patch: how many household are there nearby in the potential new spot
Similar-nearby-opt-fract		Optional new patch: fraction of similar nearby
Ethnicity	Numeric	Belonging to a specific group
My-%-similar-wanted	Numeric	The threshold for this particular agent
Tolerance-level-neighbours	Numeric	Influence of tolerance to the neighborhood composition

6 Results

In the ALLBUS, people were asked in which residential area they like to live and in which residential area they would not like to live. Based on the vignettes of Farley et al. [12], a typical residential area was drawn. This consisted of 49 units with the respondents unit in the middle and 48 units around it in order to adapt it to the German context [13–15]. The proportion of foreigners was varied over the 13 vignettes. The gap between the individual vignettes was approximately 8%. This agreement with the individual vignettes and the proportion of foreigners is shown in Table 2 [13–15]. The results are shown in the table. A short analysis of the item can also be found in Friedrichs and Triemer [16]. It can be seen that homogeneous residential areas are also rejected by a small proportion. It is therefore more realistic for purely homogeneous neighbourhoods to be rejected and thus for a certain number of the other social group to live there. It turns out that many people prefer a proportion of foreigners between 8 and 50%, which are clearly different values than those assumed by Schelling. It is also evident that a majority of the participants do not want to live in neighbourhoods where the porpotion of foreigners exceeds 66%.

In the model this implications were implemented by the variables %-similar-wanted and %-unsimilar-wanted. In order to better adapt the values to reality, they were randomly selected from the range between 0 and the set limit in the sliders. To answer the research question of how the different values affect the model, an experiment was created in the Behaviorspace in Netlogo. The different values for "density", "%-similar-wanted" and "%-unsimilar-wanted" were varied. The parameter variations can be found in Table 3. These variations are orientated on the results of the

Table 2 Calculations from the ALLBUS 2006

		Want to live in the neighbourhood	Do not want to live in this neighbourhood
Vignette number	Percentage of foreigners	2006	2006
1	0	0.81	0.05
2	8.3	0.81	0.02
3	16.6	0.73	0.03
4	25	0.56	0.07
5	33.3	0.4	0.12
6	41.6	0.29	0.19
7	50	0.19	0.27
8	58.3	0.11	0.36
9	66.6	0.06	0.46
10	75	0.03	0.58
11	83	0.01	0.68
12	91.6	0.01	0.76
13	100	0.01	0.99

descriptive analysis of the data. The maximum of the slider "%-unsimilar-wanted" is set at 20% because in the next step under 50% of the people want to live in this area. The variation of the slider "%-similar-wanted" is oriented on the empirical data, because if around 75% people want to live in a neighbourhood like in vignette 3 and only 3% do not want to live there. The density varies between 80 and 95 because 80 is the value Schelling [1] choose and between 90 and 95 because most cities in Germany have a vacancy rate of 5 to 10%. 100 runs were carried out per parameter combination with a termination after 500 steps, in total there were 8000 runs of the model.

First, with the new parameters no longer result in a state of equilibrium in the model, but the runs have to be aborted. In this experiment is terminated prematurely after 500 ticks. This means that the addition of a further behavioural assumption alone ensures that the model no longer reaches an equilibrium state after a few rounds. This is an interesting observation, as it shows that Schelling's theoretical foundation alone is not sufficient. Figure 1 shows the variation of the means of the tolerance of all agents and the number of the runs. The tolerance values increase plateau-wise, which is related to the selected population density. This makes it evident that increasing the lower limit of the people in another group causes the mean values of the tolerance to rise. The number of free places on the field therefore has an influence on the tolerance of both groups. The tolerance values rise up to 60%, which means that society seems to become more tolerant simply because a small amount of mixing is desired by the people. This is particularly interesting because Schelling's model assumed that if people did not want to be in a minority position in their neighbourhood, that alone would lead to segregation. In this model, it can be seen that the desire for actors from the other group in the neighbourhood do not lead to segregation. This assumption is very simple and needs more research.

Figure 2 shows the distribution of the mean of tolerance depending on the steps and variation of the parameters and the corresponding confidence intervals. The confidence intervals are overlapping, which means that there is no clear difference between the experiments is possible. The mean is differentiating between the experiments and is lower with a density of 95 than with a density of 80. This means that the mean values for tolerance are lower with a greater density, which would mean that the more agents are in a field, the less tolerance appears to spread. This would mean that only agents in very dense settlements appear to be more tolerant and that the own-group preference is actually strengthened as a result. Another interesting

Table 3 Variation of the parameters in the experiment

Concept	Name of the parameter	Values
Random effect	Noise	0.03
Density	Density	80, 85, 90, 95
Surrounded by agents of the same social group	%-similar-wanted	50, 60, 70, 80
Surrounded by agents of the other social group	%-unsimilar-wanted	0, 5, 10, 15, 20

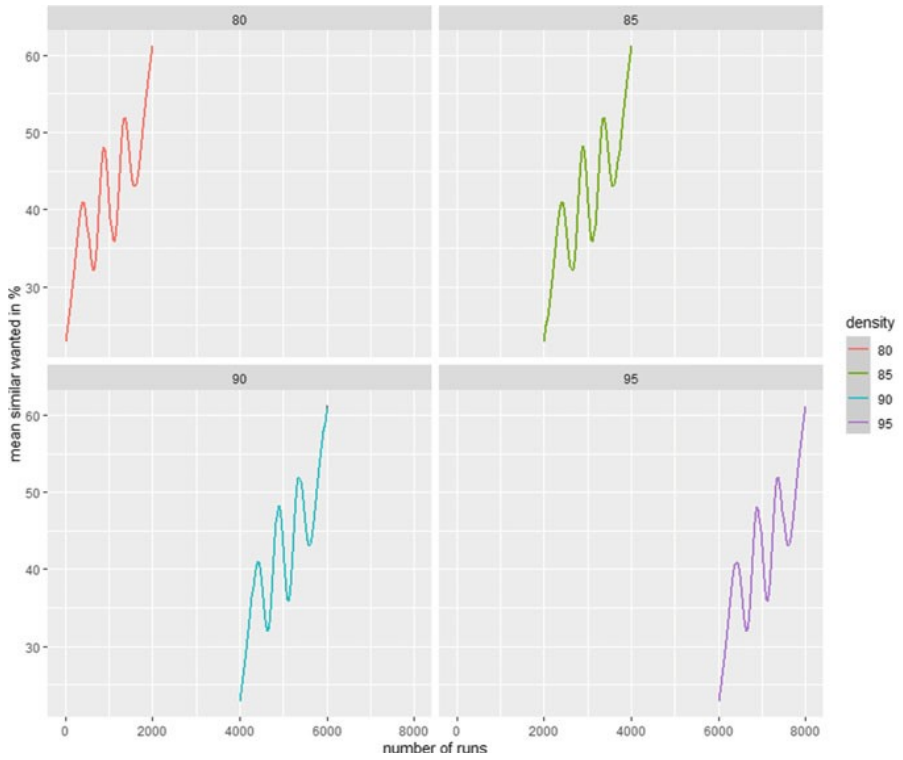


Fig. 1 Mean of the similar wanted depending on the number of runs, separated plots by density

observation is that the mean value is slightly higher at a density of 85 than at a density of 90. These effects should be considered and examined in further analyses.

Figure 3 shows the average of percent of an agent neighbors has the same colour like the agent. It can be seen that on average across all densities approx. 85% of the neighbours have the same colours as the agents. However, this also suggests that segregation is not stable, as the model must be provided with a stop condition. It is striking that the similarity is below 90%. Further analyses should examine whether this effect becomes larger or smaller with more exact values. Above all, the effect suggests that the addition of values indicating that other social groups should also be represented actually makes the neighbourhood more diverse.

7 Summary and Limitations

In this short paper, the theoretical basis for segregation has been presented first. For this purpose, the theories of Schelling [1] and Elias and Scotson [4] were linked. This provided a theoretical foundation for the agents' chains reaction. Subsequently, the

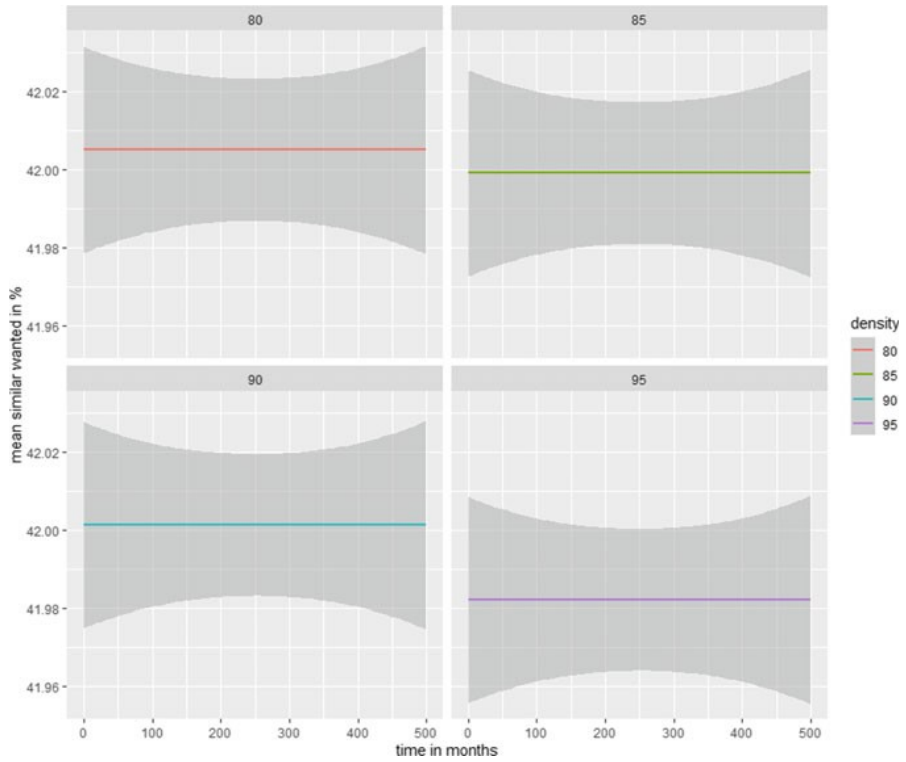


Fig. 2 Distribution of mean of tolerance depending on the steps and variation of the parameters and corresponding confidence interval

model developed was presented in Netlogo and its mode of operation was described. The rules followed the theoretical foundation were translated in mathematical terms and implemented in the model. With the ALLBUS results, individual preferences about ethnic composition can be fed into the segregation model in a more complex but also empirically more realistic way. In this paper, an agent-based model is presented that works with the empirical preferences from the ALLBUS. The distributions are rudimental implemented in the model. It turned out that the different preferences ensured that the agents became more tolerant overall and accepted more agents of the other group in their environment over time. Looking at the models also shows that the respective segregated areas in the model do not remain stable as in Schelling's model, but move. These mechanisms need to be investigated further. In addition, the values from the ALLBUS should be implemented more precisely in further models. At present, only random numbers are implemented as lower limits to implement a level of uncertainty as well. In later models, the empirical distributions are to be implemented as well as lower and upper limits. At the moment the individual tolerance is set at the beginning and not changing anymore. The individual tolerance can be shaped as well by the corresponding neighbourhood. This should be implemented

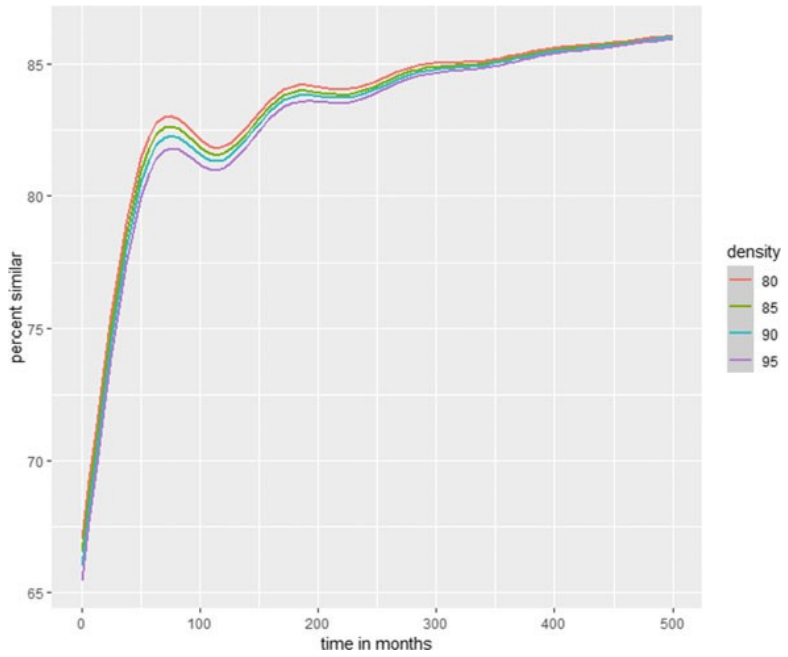


Fig. 3 Average of percent of an agent neighbors has the same colour like the agent

in further models. It is important to mention that the behavioural rule from Schelling is implemented in this model, which is critical and should be investigated in future research, if this rule correctly reflects the real world. In addition, only a simple descriptive evaluation was carried out, with the extension of the agents' attributes, better cause-effect relationships can be established. It would be interesting to investigate how the models behave when there are four groups of agents with different distributions. There is still more possibilities to use empirical data for agent-based models. In the next step more attributes should be added to the agents and as well attributes should be added to the patches.

References

1. Schelling, T.C.: Micromotives and macrobehavior. [New ed.] with a new preface and the Nobel Lecture. Norton (Fels lectures on public policy analysis), New York, NY (2006)
2. Candipan, J., Phillips, N.E., Sampson, R.J., Small, M.: From residence to movement: the nature of racial segregation in everyday urban mobility. Urban Stud. **58**(15), 3095–3117 (2021). https://doi.org/10.1177/0042098020978965
3. Hatna, E., Benenson, I.: The Schelling model of ethnic residential dynamics: beyond the integrated—segregated dichotomy of patterns. JASSS **15**(1). https://doi.org/10.18564/jasss.1873

4. Elias, N., Scottson J.L.: (1993): Etablierte und Außenseiter. Zur Theorie von Etablierten-Außenseiter-Beziehungen
5. Anderson, T., Leung, A., Dragicevic, S., Perez, L.: Modeling the geospatial dynamics of residential segregation in three Canadian cities: an agent-based approach. Trans. GIS **25**, 948–967 (2021). https://doi.org/10.1111/tgis.12712
6. Liu, Z., Li, X., Khojandi, A., Lazarova-Molnar, S.: On the extension of Schelling's segregation model. In: 2019 Winter Simulation Conference (WSC), pp. 285–296 (2019). https://doi.org/10.1109/WSC40007.2019.9004848
7. Zuccotti, C., Lorenz, J., Paolillo, R., Sánchez, A., Rodríguez, S., Selamavit, S.: Exploring the dynamics of neighborhood ethnic segregation with agent-based modelling. Empiric Appl Bradford (2021). https://doi.org/10.31235/osf.io/gmzdp
8. Flache, A., de Matos Fernandes, C.A.: Agent-based computational models. In: Gianluca Manzo (Hg.): Research handbook on analytical sociology. Cheltenham, UK and Northhampton MA, USA: Edward Elgar Publishing (Research Handbooks in Sociology), S. 453–473 (2021)
9. Farley, R., Schuman, H., Bianchi, S., Colasanto, D., Hatchett, S.: Chocolate city, vanilla suburbs:" Will the trend toward racially separate communities continue? Soc. Sci. Res. **7**(4), S. 319–344. https://doi.org/10.1016/0049-089X(78)90017-0
10. Wilensky, U.: NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (1999)
11. Wilensky, U.: NetLogo Segregation model. http://ccl.northwestern.edu/netlogo/models/Segregation. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (1997)
12. Farley, R., Frey, W.H.: Changes in the segregation of whites from blacks during the1980s: small steps toward a more integrated society. Am. Sociol. Rev. **59**(February), 23–45 (1994)
13. GESIS—Leibniz-Institut für Sozialwissenschaften.: Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2006. ZA4500 Datenfile Version 2.0.0. GESIS Datenarchiv, Köln (2011)
14. GESIS—Leibniz-Institut für Sozialwissenschaften.: Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2016. ZA5250 Datenfile Version 2.1.0. GESIS Datenarchiv, Köln (2017)
15. Wasmer, M., Scholz, E., Blohm, M.: ZUMA-Methodenbericht 2007/09. Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2006. Hg. v. GESIS—ZUMA (2007)
16. Friedrichs, J., Triemer, S.: Gespaltene Städte? Soziale und ethnische Segregation in deutschen Großstädten, 2nd edn. VS Verlag für Sozialwissenschaften/GWV Fachverlage GmbH Wiesbaden, Wiesbaden (2009)

Agent Decision-Making Heterogeneity—Agent (Meta)Frameworks for Agent-Based Modelling

Harko Verhagen, Corinna Elsenbroich, and Nanda Wijermans

1 Introduction

Agent-based models (ABM) serve amongst others to understand how social or macro phenomena can result from individual interactions with other agents and an environment as well as how these macro phenomena feed back to the individual agents. Starting off with implementing simple reactive behaviours in cellular automata (reacting to what a neighbouring cell does), ABM has increasingly moved towards modelling more cognitively complex agents [2]. Binary choices become sets of choices, criteria for choice selection have to be found, different levels of social engagement need to be modelled. It means that the agent behaviours in an ABM become more realistic (i.e. to be more like real people behave, based on concepts, models, and theories of human behaviour) also means that models become less tractable. Competing paradigms between KISS (Keep it simple stupid!) [1] and KIDS (Keep it descriptive stupid!) [7] have been battling it out as has the debate about validating ABM, in particular about validating the rules going into the model e.g. [27]. The call for realism in agent modelling has been part of the history of the field of social simulation, at least for the non-technical oriented part of the modelling community, and is presented in a clear way in EROS (Enhancing the realism of simulation) [13]—although the paper has a bias towards implementing psychological theories, reserving social

H. Verhagen (✉)
Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden
e-mail: verhagen@dsv.su.se

C. Elsenbroich
University of Glasgow, Glasgow, UK
e-mail: Corinna.Elsenbroich@glasgow.ac.uk

N. Wijermans
Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden
e-mail: nanda.wijermans@su.se

Institute for Future Studies, Stockholm, Sweden

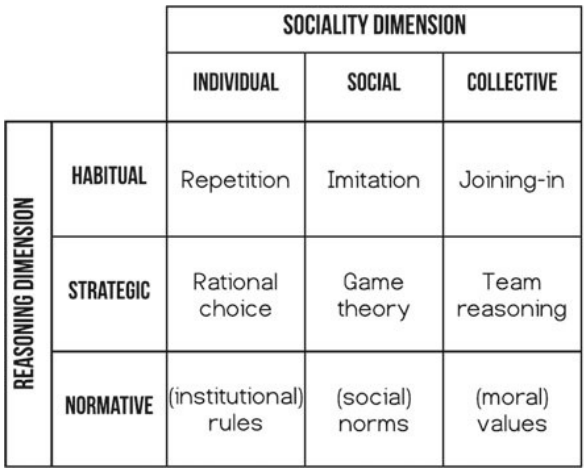
C. Elsenbroich and H. Verhagen (eds.), *Advances in Social Simulation*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-031-57785-7_48

theories to be seen as the resulting macro effects of the micro interactions, cutting the micro-macro-micro linkages in half.

One of the developments within KIDS is the recognition of decision heterogeneity between agents as well as for the same agent at different times within the simulation. Inter-agent heterogeneity is already part of simple game theoretic models with populations of mixed strategies, e.g., Hawks and Doves [16]. In cognitive models, these differences in behaviour and decision making are often encoded as differences in an agent's internal states (e.g., beliefs, norms, utility), as is the implied target of EROS [13]. For example, Fearlus is a goal driven multi-dimensional utility architecture to model common pool resource management [5, 19]. The agents decide what to do based on one of three possible strategies (statisficing, imitation, or innovation) depending on their performance and individual characteristics/preferences. The ASSOCC model [6] integrates needs, goals, and social norms to model the behaviours of agents during the COVID-19 pandemic. The different factors are integrated via modelling each of them as needs and a weighing mechanisms expressing need preferences. The Consumat is a context driven architecture with agents using different decision mechanisms in different situations [14, 15]. The context dependent decision making is developed using dimensions of *cognitive cost* and *(un)certainty* in a 2 by 2 matrix, connected to psychological theories. The MoHuB framework aggregates a huge variety of cognitive architectures into a generalised meta-framework to enable finding, communicating and eventually integrating theories from the social and behavioural sciences in models, thereby thus reflecting context sensitivity by explicitly distinguishing between what an agent/human knows and what becomes accessible/activate in a given context [21]. Finally, the Model Social Agent is an analytical framework grounded in cognitive and social science theories and concepts [4]. It consists of a 5 by 6 matrix of the types of knowledge and the cognitive processing capabilities respectively that are needed to allow for different types of social behaviour, the authors apply this matrix to two theories familiar to and used in ABM, namely Festinger's Social Comparison Theory [10] and Turner's Social Interaction Theory [23], to illustrate how these theories map onto different knowledge and processing combinations in the framework and thus put demands on how to model agents for these.

The Contextual Action Framework for Computational Agents (CAFCA) [9] a conceptual framework for context sensitive decision making, with a focus on distinguishing two levels of sociality - taking others into account as acting entities and being part of a social system that steers ones decision-making, see Fig. 1.

Like MoHuB, CAFCA is a meta-framework but different from MoHuB, CAFCA does not try to generalise over the internal processes of an agent's decision making but rather classifies the contexts in which decisions can occur, or contexts which can make a difference to decisions. It is starting out with two dimensions that are most prevalent in agent-based modelling, the social setting of a decision and whether the situation is habitual, strategic, or normative. CAFCA is first and foremost a framework for a modeller to think about which decision mechanisms might need to be incorporated into a model. Only secondly should it be used to inform the internal dynamics of agent decision making.



		SOCIALITY DIMENSION		
		INDIVIDUAL	SOCIAL	COLLECTIVE
REASONING DIMENSION	HABITUAL	Repetition	Imitation	Joining-in
	STRATEGIC	Rational choice	Game theory	Team reasoning
	NORMATIVE	(institutional) rules	(social) norms	(moral) values

Fig. 1 Contextual Action Framework for Computational Agents (CAFCA) applied to common theories. Adapted version of [9]

CACFA moves the complexity of agent decision making into the recognition of the context instead of it being a multi-dimensional architecture of (a subset of) needs, values, emotions, relationships, and utilities such as the examples described above. The purpose of CAFCA is for the modeller to think carefully which decision contexts are needed in a particular simulation model. For example, early game theoretic agent-based models (social/strategic) were extended by social norms, adding the CAFCA dimension of social/normative to the model, e.g. [3] In [8] the collective dimension of team reasoning is added to a standard game theory question of common pool problems.

CAFCA does not prescribe how these different dimensions are implemented but opens new ways of thinking about decision making in agent-based models. It thus widens the social ontology considered in ABM (Sect. 2). It also informs new ways for data collection for ABM, as can be seen in an example of its application to fisheries (Sect. 3).

2 CAFCA and Social Ontology

Social ontology is the part of philosophy analyzing the nature and properties of the social world. "What exists" is a contentious part of the social sciences, ranging from positions in which only individuals exist (individualism) over methodological positions that explanations in the social sciences need to start from the individual (methodological individualism) to positions which ascribe existence and causal powers to extra-individual entities, such as institutions, norms and social groups.[1]

ABM has often been associated with methodological individualism as a research paradigm [18]. Whilst of course the individual agent and an "agent-centric worldview" is a central tenet of ABM, it is not limited to an ontology of individuals [28].

[1] The debate is also known as the agent-structure debate, for more detail see [12].

CAFCA contributes to the ontology of agent-based modelling by explicitly drawing attention to two types of sociality, the dimension of the social, focusing on social interaction and the dimension of the collective, focusing on social belonging/inclusion. CAFCA's social dimension starts with the individual agent level, where no other agents (are seen to) exist or influence the decision-making of the agent. Context in which the presence of other agents in the environment, and their actions and decisions influence the agent's decision-making forms the social level. The final category is the collective level, building on the work in social ontology on collective intentionality and group mind. Here, agency is at the collective (team, group) level rather than the individual agent level [11]. Thus, other agents in the same collective are seen as part of the collective decision-making rather than as individual decision-making agents with which one interacts. Below we will see how this makes a difference.

3 CAFCA Research Design and the Need for Analysis of Data

As described in [22], for accurate modelling of the COVID epidemic, fine-grained data is needed to represent specific local conditions and the social reactions of individuals. Most COVID models are aiming for adequate estimates of disease spreading at a macro level but fail to cover the relevant behavioural and social complexity of societies under pandemic crisis. This limited their usability for meso and micro level analyses and predictions. Another example of the need for more detailed and fine-grained data concerns research on common-pool research problems. In most of the research, individuals are modelled as rational choice type of decision makers, where the social level is thus limited to CAFCA's social level while claiming to investigate collective consequences and behaviour. Moreover, in experimental settings to investigate more detailed interaction processes, the number of participants is small, usually not known to each other, and communication is not included, effectively excluding any collective level phenomena. In the following two sections we will look closer at these two examples.

3.1 Example of CAFCA Data (Un)Availability: The Coronavirus and the Social Impacts on Great Britain *Dataset*

The Office for National Statistics (ONS) in the UK started a large weekly survey at the end of April 2020 collecting data concerning the COVID epidemic.[2] People were asked about their behaviours related to reducing the spread of COVID-19,

[2] The ONS Coronavirus and the social impacts on Great Britain Dataset can be found at https://shorturl.at/fmoL1.

including wearing face coverings, social distancing, avoiding physical contact, and self isolation. People were also asked whether they were worried about COVID.

In Fig. 2 we can see that adherence to behaviours to reduce the spread of COVID-19 change over time. The red line shows the proportion of the population worried about COVID-19 which is consistently high, around 75% until the end of 2020, then dropping to about half the population. Whilst wearing a face covering remains high, social distancing and the avoidance of physical contact reduce in line with the reduction of people being worried about COVID.

With CAFCA we can start to interpret this as the behaviours coming from different motivations. Face coverings, social distancing, and physical contact are highly visible behaviours, likely to result in imitation or the establishing of a social norm. Why then is the adherence to face coverings high and the social interaction ones reducing rapidly? One interpretation is that wearing a face covering is a relatively easy behaviour to maintain, leading to consistently high levels through imitation. In contrast, social distancing and avoiding physical contact can be seen as difficult behaviours to maintain. Observing decreasing levels of social distancing and avoidance of physical contact can lead to a feedback with more and more people increasing social contacts. Unfortunately, the survey did not ask participants for their reasons for maintaining or changing those behaviours.

The survey did however ask about different reasons for self isolation. Figure 3 shows a range of reasons for self isolation. They range from self isolating due to worry or advise to isolate due to medical treatment (rational choice or team reasoning), over being told to self isolate (rule or norm following), to reacting to having been in contact with someone who has tested positive for COVID-19 (social norm or even a moral value expression).

CAFCA can help understand behaviours and categorise motivations. If questions about motivations were included in surveys as standard, e.g. questions why people decided to no longer socially distance or avoid physical contact, we might have a better understanding on why these behaviours changed so much faster than the use of face coverings.

3.2 CAFCA and Common-Pool Resource Problems

In a project concerning the study of collective action in common-pool resources (CPR), see [24], we have used CAFCA in several ways and noticed several challenges concerning data collection, e.g. [25, 26].

Initially we used CAFCA to take a step back and reflect on what decision-making modes are used in an agent-based model of a common pool resource dilemma [26]. This first encounter centred around the sociality dimension of CAFCA and resulted in the wish to extend the agent model to reflect the ‘collective’ CAFCA mode. This triggered an analysis of the available data on the notion of ‘togetherness’ to elicit hints of how such a formalised collective mode could manifest itself.

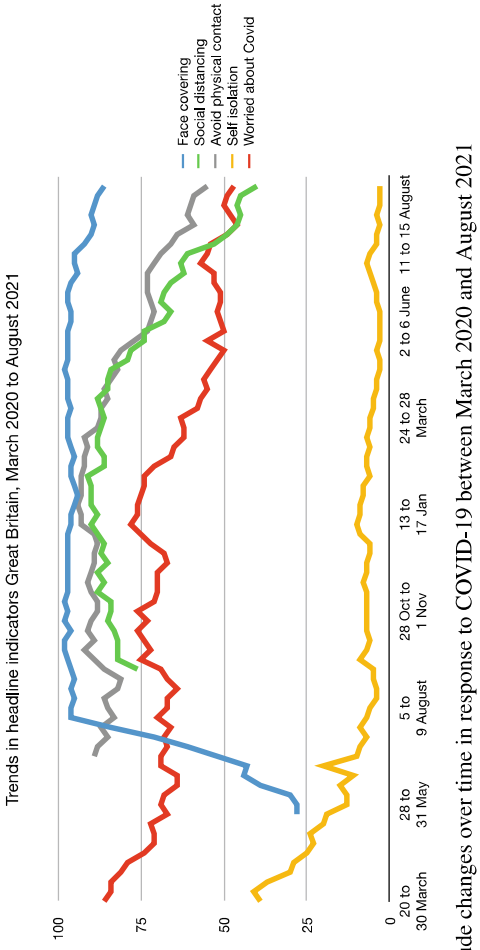


Fig. 2 Behaviour and attitude changes over time in response to COVID-19 between March 2020 and August 2021

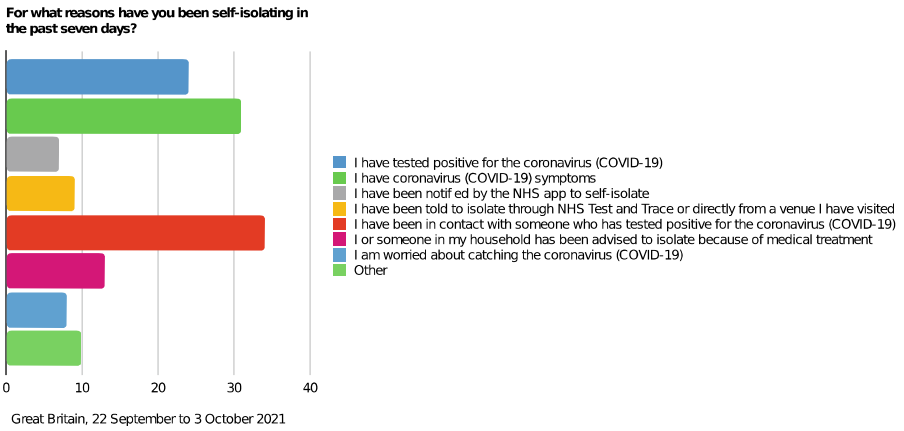


Fig. 3 Reasons for self isolation

This first interaction with CAFCA resulted in meaningful reflections on how our agents are equipped to act as social versus collective. This in turn also implied we needed to rethink how decision situations ask for agent heterogeneity in terms of how decisions are made depending on the context. This consequently also calls for data on the context in which decisions are made and how agents express their take on the context. Leading to a focus on the reasoning dimension of CAFCA.

Behaviour experiments are a common approach toward a more generic understanding of cooperative and sustainable behaviour in CPRs [17, 20]. These studies are oriented to group behavioural patterns rather than individual decision-making and focus on quantitative data. However, qualitative data would be needed to investigate the details of the decision-making modes at play and the transitions between them for a CAFCA-esque model to go beyond most CPR simulation models. Indeed, most modellers turn to different theories from the social and behavioural sciences. However, most of these theories concern relatively simple decisions with few alternatives, short evaluation horizons, and tend to be within their own domain silo.

Embedded in a series of studies on resource user decision-making, we connect to the field research in fishing communities in Thailand and Colombia—integrated in ongoing studies with other researchers. The aim of the project for to formalise the influence of perception of change in the resource on the participants' actions via internal characteristics and processes, see [24] for a detailed description of the studies. In these studies, the data gathering consists of:

1. the resource extractions/harvesting done by participants in a common-resource problem game experiment (rounds of deciding how much fish to catch) in a small group of 4 local fishers;
2. observational data of the communication between the experiment participants;
3. questionnaire data with the experiment participants; and
4. interviews with the observers (semi-structured, audio recorded and transcribed).

In addition to the adaptations of the basic approach, we exploratory analysed existing data and adapted the data collection to study the reasoning dimension. Our re-analysis from the CAFCA reasoning dimensions demonstrates how difficult it was to elicit hints to what mode people are in when making decisions [25]. We therefore included a question about the argumentation style in the observation scheme and in the end interview with the experiment team members.

The observation protocols and interviews did provide some detailed data. A preliminary analysis of the data shows that the relatively small number of interactions (the experiments consist of 20 rounds of decision-making on how much fish to catch) does not offer many opportunities for new social effects that emerge during the experiment. Rather, the pre-existing knowledge of the other participants (part of the same fishing community) plays a role as well as general socio-cultural norms (e.g., one should respect the elderly, thus their voice in the decision-making process has extra weight) [24].

The need for detailed information puts high demands on the observers. As the experiments were carried out in countries of which the researchers did not speak the local language(s). Thus, the observers need to be able to express themselves clearly in English. As we are unaware of all culture-related details while the data comes to us via our interviews of the observers, the observers need to have enough distance to their own cultural background to recognize the interesting processes and utterances from the experiments. To note any transitions between individual, social, and collective they need to observe what is obvious and therefore invisible to the insiders .

4 Discussion and Conclusion

Agent-based modelling needs to constantly walk the tightrope between realism and tractability. We want our models to be simple but no simpler than the problem or research question demands. We hope CAFCA can help to find the right level of abstraction for a model by helping to think through which decision dimensions are relevant in the particular setting. For example the Consumat uses individual and social habitual and strategic dimensions of decision making to replicate consumer behaviours. A model of mask wearing might need a population of collective normative agents, some social and collective habitual ones and some social strategic ones to replicate the dynamics of adherence to COVID restrictions.

In addition to allowing the modeller to think about the dimensions needed to model a given target, CAFCA can then also help with the appropriate data collections by focusing surveys, interviews, and focus groups on why people do things and how they might change their behaviour depending on contexts. This may include the need for training of those that help modellers collect the data needed.

Acknowledgements Elsenbroich was supported by the MRC/CSO Social and Public Health Sciences Unit core grant (MR/ S037578/1 and MC_UU_00022/5, and Scotland Chief Scientist Office

Grant SPHSU 20) and the UK Prevention Research Partnership (MR/S037578/1). Wijermans was supported by the Swedish Research Council Formas [grant 2018-00401].

References

1. Axelrod, R.: On six advances in cooperation theory. Analyse Kritik **22**(1), 130–151 (2000)
2. Balke, T., Gilbert, N.: How do agents make decisions? a survey. J. Artif. Soc. Soc. Simul. **17**(4), 13 (2014). https://doi.org/10.18564/jasss.2687, http://jasss.soc.surrey.ac.uk/17/4/13.html
3. Bicchieri, C.: The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press (2006)
4. Carley, K., Newell, A.: The nature of the social agent. J. Math. Sociol. **19**(4), 221–262 (1994)
5. Cioffi-Revilla, C., Gotts, N.: Comparative analysis of agent-based social simulations: Geosim and fearlus models. J. Artif. Soc. Soc. Simul. **6**(4) (2003)
6. Dignum, F.: Social Simulation for a Crisis. Springer (2021)
7. Edmonds, B., Moss, S.: From kiss to kids–an 'anti-simplistic' modelling approach. In: Multi-Agent and Multi-Agent-Based Simulation: Joint Workshop MABS 2004, Revised Selected Papers 5. pp. 130–144. Springer, New York (2005)
8. Elsenbroich, C., Payette, N.: Choosing to cooperate: modelling public goods games with team reasoning. J. Choice Model **34**, 100203 (2020)
9. Elsenbroich, C., Verhagen, H.: The simplicity of complex agents: a contextual action framework for computational agents. Mind Soc. **15**(1), 131–143 (2016)
10. Festinger, L.: A theory of social comparison processes. Hum. Relat. **7**(2), 117–140 (1954)
11. Hakli, R., Miller, K., Tuomela, R.: Two kinds of we-reasoning. Econ. Philos. **26**(3), 291–320 (2010)
12. Hollis, M., Smith, S.: Two stories about structure and agency. Rev. Int. Stud. **20**(3), 241–251 (1994)
13. Jager, W.: Enhancing the realism of simulation (eros): on implementing and developing psychological theory in social simulation. J. Artif. Soc. Soc. Simul. **20**(3) (2017)
14. Jager, W., Janssen, M.: An updated conceptual framework for integrated modeling of human decision making: the consumat II. In: Paper for Workshop Complexity in the Real World@ ECCS, pp. 1–18 (2012)
15. Jager, W., Janssen, M.A., Viek, C.: Experimentation with household dynamics: the consumat approach. Int. J. Sustain. Develop. **4**(1), 90–100 (2001)
16. Klemens, B.: Finding optimal agent-based models. Tech. Rep. **49** (2007)
17. Lindahl, T., Janssen, M.A., Schill, C.: Controlled behavioural experiments. In: Biggs, R., Vos, A.D., Preiser, R., Clements, H., Maciejewski, K., Schlüter, M. (eds.) The Routledge Handbook of Research Methods for Social-Ecological Systems, pp. 295–306. Routledge, London (2021). https://doi.org/10.4324/9781003021339-25
18. Manzo, G.: Agent-based models and methodological individualism: are they fundamentally linked? L'Annee sociologique **70**(1), 197–229 (2020)
19. Polhill, J.G., Sutherland, L.A., Gotts, N.M.: Using qualitative evidence to enhance an agent-based modelling system for studying land use change. J. Artif. Soc. Soc. Simul. **13**(2), 10 (2010)
20. Poteete, A.R., Janssen, M.A., Ostrom, E.: Working Together: Collective Action, the Commons, and Multiple Methods in Practice. Collective Action, the Commons, and Multiple Methods in Practice. Princeton University Press, Princeton (2010)
21. Schlüter, M., Baeza, A., Dressler, G., Frank, K., Groeneveld, J., Jager, W., Janssen, M.A., McAllister, R.R.J., Müller, B., Orach, K., Schwarz, N., Wijermans, N.: A framework for mapping and comparing behavioural theories in models of social-ecological systems. Ecol. Econ. **131**, 21 – 35 (2017). https://doi.org/10.1016/j.ecolecon.2016.08.008

22. Squazzoni, F., Polhill, J.G., Edmonds, B., Ahrweiler, P., Antosz, P., Scholz, G., Chappin, E., Borit, M., Verhagen, H., Giardini, F., Gilbert, N.: Computational models that matter during a global pandemic outbreak: a call to action. J. Artif. Soc. Soc. Simul. **23**(2) (2020). https://www.jasss.org/23/2/10.htm
23. Turner, J.H.: A Theory of Social Interaction. Stanford University Press (1988)
24. Wijermans, N., Schill, C., Lindahl, T., Schlüter, M.: Combining approaches: looking behind the scenes of integrating multiple types of evidence from controlled behavioural experiments through agent-based modelling. Int. J. Soc. Res. Methodol. 1–13 (2022). https://doi.org/10.1080/13645579.2022.2050120
25. Wijermans, N., Verhagen, H.: Formalising agent reasoning—the Paso Doble of data and theory. In: Proceedings of the Social Simulation Conference (2022)
26. Wijermans, N., Verhagen, H.: Fishing together?—exploring the murky waters of sociality. In: Dam, K.H.V., Verstaevel, N. (eds.) Multi-Agent-Based Simulation XXII—22nd International Workshop, MABS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers. Lecture Notes in Computer Science, vol. 13128, pp. 180–193. Springer (2021). https://doi.org/10.1007/978-3-030-94548-0_14
27. Windrum, P., Fagiolo, G., Moneta, A.: Empirical validation of agent-based models: alternatives and prospects. J. Artif. Soc. Soc. Simul. **10**(2), 8 (2007)
28. Zahle, J., Kincaid, H.: Agent-based modeling with and without methodological individualism. In: Advances in Social Simulation: Looking in the Mirror, pp. 15–25. Springer (2020)