Simulating Human Routines

Integrating Social Practice Theory in Agent-Based Models

Mercuur, R.A.

DOI
10.4233/uuid:4b70aa0a-8c13-421d-9043-6274311df2aa

Publication date
2021

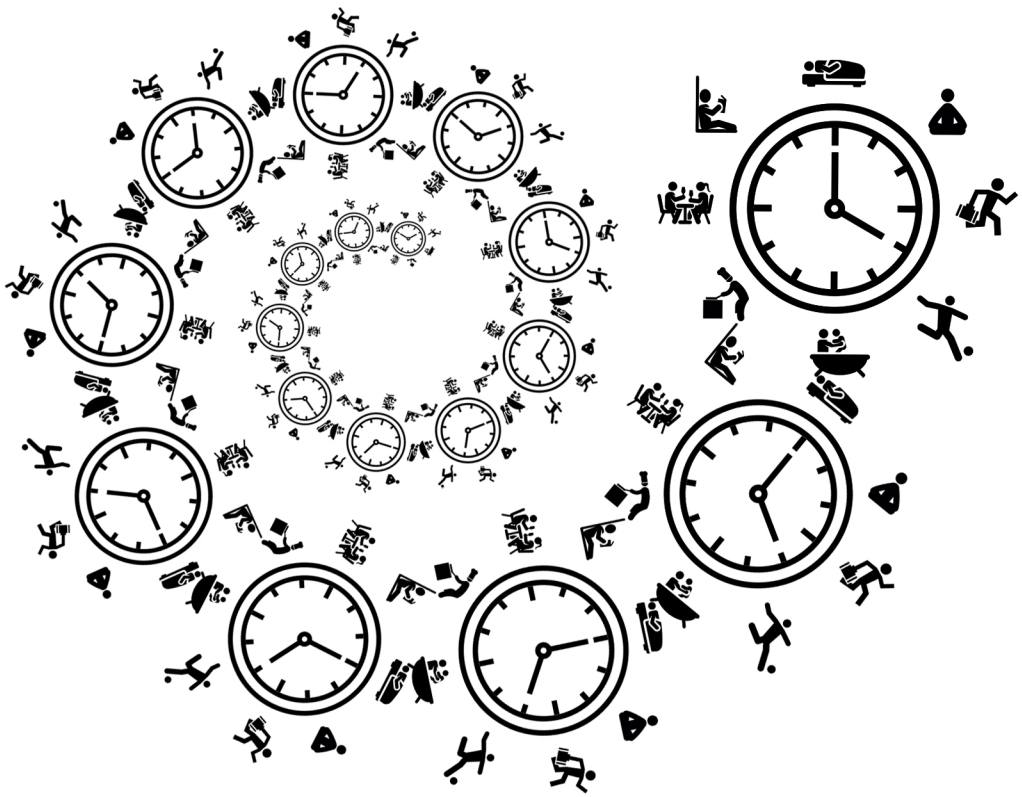
Document Version
Final published version

Citation (APA)
Mercuur, R. A. (2021). *Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:4b70aa0a-8c13-421d-9043-6274311df2aa

SIMULATING HUMAN ROUTINES

Integrating Social Practice Theory
in Agent-Based Models

RIJK MERCUUR



Simulating Human Routines

Integrating Social Practice Theory
in Agent-Based Models

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Wednesday 26th of May 2021 at 10:00 o' clock

by

Rijklof Aschwin MERCUUR

Master of Science in Artificial Intelligence,
Universiteit van Utrecht, Utrecht, the Netherlands born in Nijmegen, the
Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	voorzitter
Prof. dr. C.M. Jonker	Technische Universiteit Delft
Dr. M.V. Dignum	Technische Universiteit Delft
Independent members:	
Prof. dr. D. Helbing	Technische Universiteit Delft
Prof. dr. B. Edmonds	Manchester Metropolitan University
Dr. J.G. Pollhill	The James Hutton Institute
Prof. dr. ir. G.J. Hofstede	Universiteit van Wageningen
Prof. dr. ir. M.F.W.H.A. Janssen	Technische Universiteit Delft, reserve member

Keywords: Social Practice Theory, Agent Architecture, Agent-Based Models, Simulation, Habits, Social Agents, Human values

SIKS Dissertation Series No. 2021-02. The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

This research was supported by the Engineering Social Technologies for a Responsible Digital Future project at TU Delft and ETH Zurich.

ISBN 978-94-6384-213-6

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Adopt what is useful, reject what is useless, and add what is specifically your own. Man, the living creature, the creating individual, is always more than any established style or system

Bruce Lee

Contents

Preface

The reason I love watching sports — especially team e-sports — is to see the social practice at work in full perfection. In split seconds an expert uses knowledge that is trained, automated and embodied to such a degree that one — in the flow — makes the right decision. On top of that a trusty teammate knows you will make the right call and times that pass, sets up that heal, flanks the enemy. No need to communicate. No need to explicitly know what you are doing. Just one massive tuning in; to the flow, tao, life, magic. Sports demonstrate the ability of humans to train, change and shape, and eventually let go into beautiful unified patterns of behaviour. This ability is what defines us as humans and explains the success of our species.

The discussions after an e-sport match are fascinating as well. Commentators will discuss the current view on how the game should be played, or as it is called in e-sports: the meta. What moves are strong at this moment? What moves are weak? The meta corresponds with what I'll call the collective view on a practice. Knowing the collective view in sports is an advantage because it allows a player to adapt his or her playing style. A player copies the moves that are strong or predicts what others will do and counters it. These adaptions, in turn, change how the collective views the game: changing the meta. Which in turn changes a player's decision-making... changing the meta again, and so on, and so on... These dynamics between the individual and the collective (or micro and macro) are fascinating and relate to another core principle of being human: for me I'm special, but I also know, I'm one of many.

The PhD is a discovery to find your own social practice of doing research. Successful evolution of a system depends on both copying the collective and adapting it to make it your own. I've learned about the collective view on research: the grinding, the focus on details, the grasping for true knowledge and the individual search for acknowledgement. I've tried things out and adopted what I see fit. I've added my own aspects and eventually, never intending so, created my own practice. I am far from completing such a journey but do know that my view on good research includes this magic of a seamless unified carrying out. Research should be cooperation where the creation is the product of the interaction of a group that knows themselves and creates something no individual alone can create. A process where our behaviour flows together because we cherish the same values and know what to expect of each other. Hopefully, one day, my view will too influence the meta.

Rijk Mercuur

I

Introduction

1

Introduction

Effort is not gritting once teeth and forcing action. Effort is creating the right conditions. [...] Effort is the self-confidence to overcome. Effort is joy, relaxation, enthusiasm. Finding joy in what is good.

Stephen Batchelor, Secular Buddhism

If I feel completely muddled it isn't that there's a problem that I have to solve: I just don't know who I am in connection with that problem.

Joko Beck, Everyday Zen

Parts of this chapter have been published as Mercuur, R. (2018). Realistic Agents with Social Practices. Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, 1752–1754 [168].

1

1.1. Rationale

1.1.1. Motivation and Challenge

The behaviour of groups of people plays an important but complex role in a wide range of social challenges such as disease outbreaks, climate change and coördinating a hospital. For example, in a disease outbreak, it is important to know where people go and to what extent they conform to social distancing. However, this is not easy to find out: individuals act differently, influence each other and adapt their behaviour over time. In addition, these individuals produce emergent effects that feedback on their decisions (e.g., buying toilet paper, leads to the toilet paper being out of stock, which in turn, leads to a daily shop run to find toilet paper, etc.). These complex aspects of human behaviour – heterogeneity, interactions, emergence, feedback loops and individual learning – add to the complexity of social challenges. Understanding these complex aspects of human behaviour helps us in dealing with social challenges.

Agent-based social simulation (ABSS) enables researchers to simulate these complex aspects of human behaviour by directly representing individual entities and their interactions [107] (Box 1.1.1 provides an introduction on ABSS). ABSSs are useful for a wide range of applications and a wide range of purposes. ABSSs have been applied to study opinion dynamics [62], consumer behaviour [135], industrial networks [5, 30], supply chain management [246] and electricity markets [18, 113, 216]. An intuitive purpose of ABSSs is to predict which policies will provide the sought after results. And indeed, ABSSs have been used to guide policies aimed at decreasing air pollution [99], to test vaccination strategies that guard against pandemics [244], and to quantify risks from flooding [79]. In addition to predicting and guiding policies, ABSS is used for theoretical exposition, explanation, description and illustration [80].[1] A common factor in all these ABSS studies is that they provide and communicate insights into complex social systems.

ABSS studies use different agent frameworks that emphasise different aspects of human decision-making.[2] Relatively simple agent frameworks are used in ABSS studies such as the Schelling segregation study [228]. In contrast, this thesis focuses on cognitive agent models (see Box 1.1.1). Cognitive agents, such as reasoning agents and normative agents, are based on work from the multi-agent systems community and intelligent agent community. Reasoning agents emphasise autonomy, reactivity and proactivity. For example, the BDI agent (Belief, Desire, Intention agent) reasons which actions to take next based on a set of beliefs, desires and intentions [213]. Normative agents emphasise sociality by incorporating norms: standard, acceptable or permissible behaviour in a group or society [91]. For example, the EMIL-A framework [13] models how such norms emerge from society and the BOID framework [38] models how agents use these norms to come to decisions. In parallel research based on the social-psychological literature, agents are based on social-psychological literature and model mechanisms such as fear

[1] In fact, in the ABSS community the possibility of prediction, given our current knowledge of social systems, is a recurring debate.

[2] Throughout this thesis, we use 'framework' to refer to a domain-independent conceptualisation and 'model' to refer to the domain-dependent conceptualisations based on such a framework.

(e.g., Agent-0 [83]) or needs (e.g., Consumat [135]). For example, the Consumat agent makes its decisions based on theories on need satisfaction and uncertainty. Employing the right agent framework for the right system and the right purpose is paramount for the success of an ABSS study [19].

Current agent frameworks do not support researchers in simulating human routines (see Chapter 3). Routines capture that humans make habitual decisions, interconnect these habits throughout the day and use these interconnected habits as a blueprint for social interaction (see Chapter 3). Evidence from the social sciences shows the influence of routines on human decision-making (see Chapter 3). Habitual behaviour, one aspect of routines, has theoretically been studied by Aristotle [8], William James [137] and more recently Triandis [250]. To quote William James on habits: "Ninety-nine hundredths or, possibly, nine hundred and ninety-nine thousandths of our activity is purely automatic and habitual, from our rising in the morning to our lying down each night."[Ch. VIII] [136] Empirically, meta-studies in transport choices [124], food choices [215] or recycling [148] consistently show that measures of habits have a moderate to strong correlation with behaviour. In addition to this evidence speaking for the importance of the habitual aspect of routines, there is similar theoretical and empirical evidence for the importance of the social and planning aspects of routines (see Chapter 3). Multiple scientific fields argue that integrating this evidence would improve the success of their decision-making models: multi-agent systems [240], game-theory [234], economics [144, 248] and social psychology [104]. Likewise, an agent framework that integrates human routines would enable ABSS researchers to gain new insights into social systems grounded in theoretical and empirical evidence on routine behaviour.

To illustrate the relevance of these new insights (that are gained by an agent framework that integrates routines), we discuss two examples. Part III of this thesis further demonstrates the relevance of an agent framework that integrates routines with several applications and Chapter 8 discusses our contributions to the ABSS, MAS and social science community.

Exploring Vegetarian Policies In Mercuur [167], we used an agent model that integrates habits to simulate policies that motivate people to break out of their meat-eating habits. Springmann et al. [242] calculates that if the world went vegan, this would save 69.4 billion land animals per year [85], save 8 million human lives by 2050, reduce greenhouse gas emissions by two thirds and lead to healthcare-related savings and avoided climate damages of $1.5 trillion. Current policies predominantly focus on providing information on the environmental consequences of meat-eating [2]. Dining routines are well ingrained into humans and merely hearing new information does not suffice to overcome strong habits to eat meat (e.g., [233]). Using an agent model of habits, Mercuur [167] simulates different policies that focus on breaking habits: opening up new restaurants, extending the menu of current restaurants or organising a 'vegetarian week'. These simulations show that organising a vegetarian week at a location known to the agent motivates an agent to break out of its meat-eating habit. The newly acquired meat-free habit, in turn, influences the dining behaviour in their household context, creating a

domino-effect. This doctoral dissertation improves and extends the habitual agent presented in Mercuur [167] by capturing other aspects of routines (i.e., routines are not only habitual but also social and interconnected). By enabling the simulation of all aspects of routines, we further provide insights for policy makers that help them save lives, reduce greenhouse gas emissions and cut back on economic costs.

Discerning Habitual Theories In Chapter 7, we present a simulation study that uses an agent framework of human routines to compare two theories on how habits break. Inducing behavioural change requires a good understanding of how habits break. For example, how do we motivate people to break their car driving habits and take the train to mitigate CO_2-emissions? Mercuur et al. [176] identifies two theories in the psychological literature that explain how habits break: the decrease theory [156, 160] and persist theory [3, 35]. Both theories are used to explain behavioural change, but one states the original habit fades out, while the other theory states the habit persists. Using an agent framework of human routines, we simulate a scenario where agents are motivated to do *multiple* alternative actions (e.g., take the bike or take the train), instead of *one* alternative action (e.g., take the bike). We show that in this scenario, the two theories should lead to different results. By translating the simulation experiment into an empirical experiment, social scientists will be able to find out which theory is supported by evidence and subsequently improve their chances of breaking bad habits.

An agent framework that integrates routines thus enables relevant insights in social systems, such as, insights that guide policymakers in reducing our meat intake and insights that guide social scientist in improving their theories on breaking habits.

Box 1.1.1: Introduction to Agent-based Simulations

Agent-based models enable researchers to model complex human behaviour by directly representing individual entities and their interactions [107]. These individual entities are called agents. Agent-based models are computational models: a computer program represents as clearly as possible how one believes that reality operates. The computer program enables researchers to simulate an experiment (i.e., an agent-based simulation or ABS). The model is used to simulate the real world as it might be or become in a variety of circumstances. These agent-based simulations enable researchers to gain insight into complex social challenges (this research is called agent-based social simulation (ABSS)).

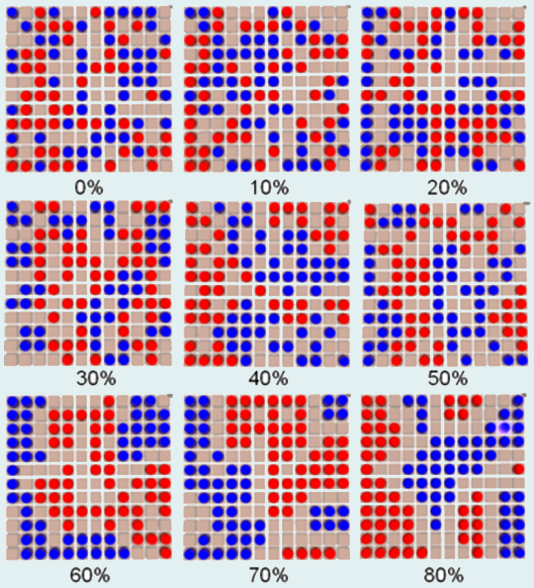


Figure 1.1: Urban areas with two kind of households (blue and red) simulated with different tolerance threshold (0%, 10%, etc.) [138]

An influential ABSS study in understanding segregation is the Schelling model [120, 228]. The model represents an urban area with two kinds of households (blue and red) (see Figure 1.1). At each timestep, each household counts the fraction of nearby households that are of the other colour. If the fraction is greater than some constant threshold, then the household considers itself to be unhappy and moves to an empty spot (grey). This relocation, in turn, influences households decisions to move in the next timestep. The surprising result of this domino-effect is that for relatively high levels of tolerance (a threshold of 30%) an initial random distribution of households segregates into 'red' neighbourhoods and 'blue' neighbourhoods. The model has been influential because it is simple, has a surprising and robust result, and gives us an interesting insight: low degrees of racial prejudice could yield strongly segregated U.S. cities [20].

1

Box 1.1.2: Simple Agents vs Cognitive Agents

This thesis uses cognitive agents – instead of relatively simple agent models – to simulate social phenomena. Simple agent models — or abstract models [105] — aim to reduce the complexity of a social system to essential (social) mechanisms to understand the dynamics and (emergent) outcomes of these mechanisms [106, 121]. For example, the Schelling model (Box 1.1.1) shows that a simple mechanism (based on tolerance) generates segregated neighbourhoods. The advantage of using simple agent models is that they are computationally feasible, easy to understand and communicate. Besides, they give surprising insights into the emergent effects caused by simple mechanisms.

Cognitive agent models trade some of these practical advantages to enable cross-validated explanations. Cross-validated explanations have been proposed in ABSS by Moss and Edmonds [188]. Simulations support the explanation of social phenomena by reproducing the phenomena via programmed mechanisms. Cross-validating an explanation implies validating three steps in such an explanatory simulation: input validation (where the input of the simulation matches the input of the system), mechanistic validation and output validation. Cognitive models enable modellers to bear evidence on the cognitive input and mechanism that produce behavioural output. Increasing the evidence that bears on models allows modellers to reduce the number of possible explanations that explain the observed behavioural output. For example, the Schelling model (see Box 1.1.1) does not explain why people become more or less tolerant nor how this tolerance interacts with other cognitive functions (e.g., imitation, habits). Thus, the Schelling model allows multiple explanations of segregation. By bearing evidence from the cognitive sciences (e.g., qualitative interviews [188]) on cognitive ABSSs, we pinpoint the most likely explanation. In short, cognitive agent models enable us to integrate evidence from the cognitive science in ABSS and pinpoint which explanations integrate this evidence.

Another advantage of cognitive agents is that they aim to model the interaction of social mechanisms and cognition [106]. Humans differ from physical particles in that they understand and intentionally influence macro-level social mechanisms that affect them. For example, voters influence the success of a party, which in turn influence the voters. Thus, humans (e.g., the party leader or voter) reason about the social level (e.g., the party) and how their actions might influence this level (and, in turn, feedback on them). An ABSS that attempts to explain the party's politics only in terms of social mechanisms (i.e., an ABSS that reduces cognitive details to social mechanisms) does not suffice. Instead, for certain use cases, ABSS researchers need to understand the dynamics of the interaction between the cognitive and social levels [106]. This thesis thus aims for socio-cognitive agent models, specifically focusing on routines, and as such supports cross-validated ABSS that model the dynamics between the cognitive and social level.

1.1.2. A Solution

An agent framework that integrates human routines would enable researchers to gain important insights into social systems. The agent framework needs to fulfil two criteria to support ABSS researchers:

Grounding in Evidence The framework needs to support models that are grounded in evidence from the social sciences. This grounding relates to a larger series of papers that call for integrating evidence in agent models [29, 80, 81, 187, 243]. Edmonds [80] points out the large number of published 'floating models' in ABSS: models that are not constrained by empirical evidence of observed social phenomena and as such do not provide insights about social phenomena. In more detail, the understanding a model provides depends on the strength of three steps in modelling: mapping into the model, inferring using the model and mapping back from the model. By utilising evidence from the social sciences, we strengthen the first step of mapping evidence on human decision-making into our model of human decision-making: agents.

Reuseable The framework needs to support reusable models: models usable by different scientists, comparable with other models and models that are easy to extend or combine with other models. Edmonds [80] argues for reusability to serve the scientific process of collective modelling. He describes ABSS as an evolutionary process where each 'generation' of models become 'fitter'. For such an evolutionary process to be successful, fit parts of previous generations, need to be compared, copied and reused. Kaminka [145] and Dignum et al. [71] argue for reusability to ensure efficiency. If modellers can build upon a domain-independent framework, they can spend their energy on problems specific to the ABSS study. In short, the reusability of a model is important for a sustainable and successful future for ABSS as this contributes to the evolution of models and efficient modelling.

To ensure the reusability of our agents, we need a domain-independent agent *framework* that enables ABSS researchers to create multiple domain-dependent agent *models*. Dignum et al. [71], Kaminka [145] and Norling [196] identify that current agent models use similar socio-cognitive theories on sociality, but cannot compare and combine their models because there is no overall framework. For example, agent models capture social mechanisms of imitation in different ways such as a neighbour-copying-function [135] or deontic logic model of norms [38]. A general framework relates these conceptualisations of norms, enabling, for example, the combination of the neighbour-copying function with a penalty for when the norm is broken specified by the deontic model. By creating a domain-independent agent framework, we enable the reuse of domain-independent concepts to ensure the reusability of our agents.

To enhance the reusability and grounding of evidence, we build the framework based on a socio-cognitive theory. Jager [134] argued for integrating socio-cognitive theories in our agent models to enhance the realism of agents. By using socio-cognitive theories we stand on the shoulders of years of social and psychological evidence. In addition, as socio-cognitive theories range over multiple domains,

they are ideal for making a domain-independent agent framework. Thus to create a domain-independent agent framework that is grounded in evidence, we need to select a socio-cognitive theory that provides the concepts, relations and evidence to model human routines.

This thesis uses social practice theory (SPT) to ground an agent framework that integrates our human routines. SPT is in particular applicable to model human routines as this theory aims to describe our 'daily doings and sayings' [227]. Our day is full of social practices (SPs): working, dining, commuting, teaching, meeting, walking or sports. SPT provides concepts and mechanisms relevant to describe the habitual nature of human routines [34]. For example, it describes that when one is at the office, one habituality enacts the SP of working. Besides, SPT provides concepts and mechanisms relevant to describe the social nature of human routines [214, 227]: a practice is not only individual but others have a similar practice. For example, SPT describes that when your colleague enters your office, he or she has a similar working practice and does not distract you but waits until the coffee break at 10:30 to discuss current matters. SPT thus helps us to conceptualise and ground the habitual and social aspects of human routines.

To create an agent framework that supports researchers in modelling human routines its necessary to integrate SPT and ABSS. Previous work has applied SPT to ABSS but needs to be extended to include the following aspects:

An integration of agent theory and SPT to ensure the framework enables modelling decision-making on human routines. Holtz [127] has made an abstract ABSS of SPT applied to consumer behaviour. Holtz [127] focuses on modelling an SP as a separate entity and does not integrate SPs with agent theory. In other words, the agents in this ABSS are the SPs themselves and not human decision-makers. However, to integrate social-psychological evidence on human decision-making we need to represent human decision-makers (i.e., agents). Thus, to integrate current evidence on human routines, we need to synthesise research from SPT, sociology, social-psychological and agent theory.

A systematic and domain-independent translation from SPT to an agent framework to ensure reusability, combinability and correctness. [191] used SPT in an ABSS that studies energy systems. Although the resulting model provides useful insights for the specific domain, without a domain-independent and systematic approach modellers lack the general context to reuse Narasimhan et al. [191]'s model for their own purposes. A domain-independent translation enables modellers to separate between domain-specific choices and reusable domain-independent choices. A systematic approach (e.g., via requirement elicitation) ensures transparency in the choices made for the framework such that ABSS researchers know which propositions they commit to when they use the framework. To ensure a transparent and systemic approach in integrating SPT in agents, we need to translate the theory to requirements, surveying a list of concepts from the literature and providing explicit model choices for each concept.

An evaluation of current agent frameworks that ensures we do not reinvent the wheel and determines the scope of the new agent framework in comparison to others. Current evaluations of agent frameworks [72, 77, 125] suggests that current agent models do not suffice to model our human routines and calls for an integration of SPT in agent models. However, without a rigorous comparison of current agent models against precise requirements, it is unclear what current models lack, for which purposes they suffice and what parts should be reused to model human routines. To ensure that the new agent framework does not reinvent the wheel, we need to evaluate current agent frameworks against SPT.

An agent framework that correctly captures SPT in computationally implementable concepts and relations. We take the high-level conceptual model by Dignum and Dignum [77] as a starting point and extend it to (1) meet the earlier determined requirements and (2) make it computationally implementable. Such a translation is not straight forward, as SPT is an ambiguous, abstract, broad, ever-expanding sociological theory that is interpreted differently by different sociologists and of which empirical evidence still has to show its exact scope [226]. In contrast, a computational model needs to be precise and unambiguous such that it is implementable in a computer language. To utilise SPT in agents, we need to translate that theory into a computational model.

Case studies testing the agent framework to ensure the applicability and scope. To our knowledge only Holtz [127] and Narasimhan et al. [191] have applied SPT in ABSS. However, they use a different translation from SPT to ABSS and therefore, do not test the applicability of our framework. In addition to testing the applicability of our framework, using SPT in ABSS studies helps us to determine the scope of SPT. For example, the discussion in Chapter 8 of this thesis explains how SPT has given interesting insights in the domain of transport mode choices but is less relevant for use cases where agents only make one decision (eg., adoption of HV-panels) or have little freedom in making their decision (e.g., the ultimatum game). Knowledge about the scope of SPT is useful for both ABSS researchers as well as a wider range of social scientists who use the theory.

1.2. Problem Statement

Given the need for an agent model that integrates human routines and the open issues identified in the previous section, we can now formulate the main objective of this thesis.

Research Objective

Create a domain-independent agent framework that integrates theories on social practices to support the simulation of human routines.

The open issues that have prevented a model that integrates social practices lead to the following key questions:

RQ1 What aspects of our human routines are (1) emphasised by social practice theory and (2) important for agent-based social simulations?

RQ2 What requirements for an agent framework that integrates these aspects represent current evidence from agent theory, social practice theory and social psychology?

RQ3 To what extent do current domain-independent agent frameworks satisfy these requirements?

RQ4 What is a domain-independent agent framework that satisfies these requirements?

RQ5 To what extent does the resulting agent framework support simulation studies on human routines?

1.3. Thesis Outline

Figure 1.2 provides an overview of the chapters and their relation with the research questions. This part contained a single chapter that introduced our aim to create an agent framework that supports the simulation of human routines. We identified that to support ABSS researchers our framework needs to integrate SPT and agent theory, correctly reflect these theories and support simulation studies.

Part II identifies the aspects of SPT that are relevant for ABSS, distils requirements from the literature, reviews current agent models and provides the SoPrA framework that satisfies said requirements. Chapter 2 provides a preliminary study that focusses on two aspects of SPs: values and norms. The chapter provides an agent framework to simulate values and norms and a comparison with the canonical homo economicus framework by simulating a psychological experiment on dividing money. By focussing only on values and norms, we make the conceptualisation of SPT more manageable. In addition, a modular modelling approach enables us to examine what each aspect of SPT offers: to compare the parts to the whole. Chapter 3 zooms out again and considers the SP as a whole. It identifies that the habitual, social and interconnected aspects of SPs are relevant for ABSS, provides a set of requirements for integrating SPT in agent models and provides an evaluation of 11 current agent models with respect to these requirements. Chapter 4 presents the Social Practice Agent (SoPrA) framework that integrates SPT and fulfils the requirements set out in Chapter 3. Part III applies SoPrA to case studies to validate the applicability and test the scope of the framework.

Part III describes three applications of SoPrA. The case studies show how our framework is able to support research in different domains with different purposes

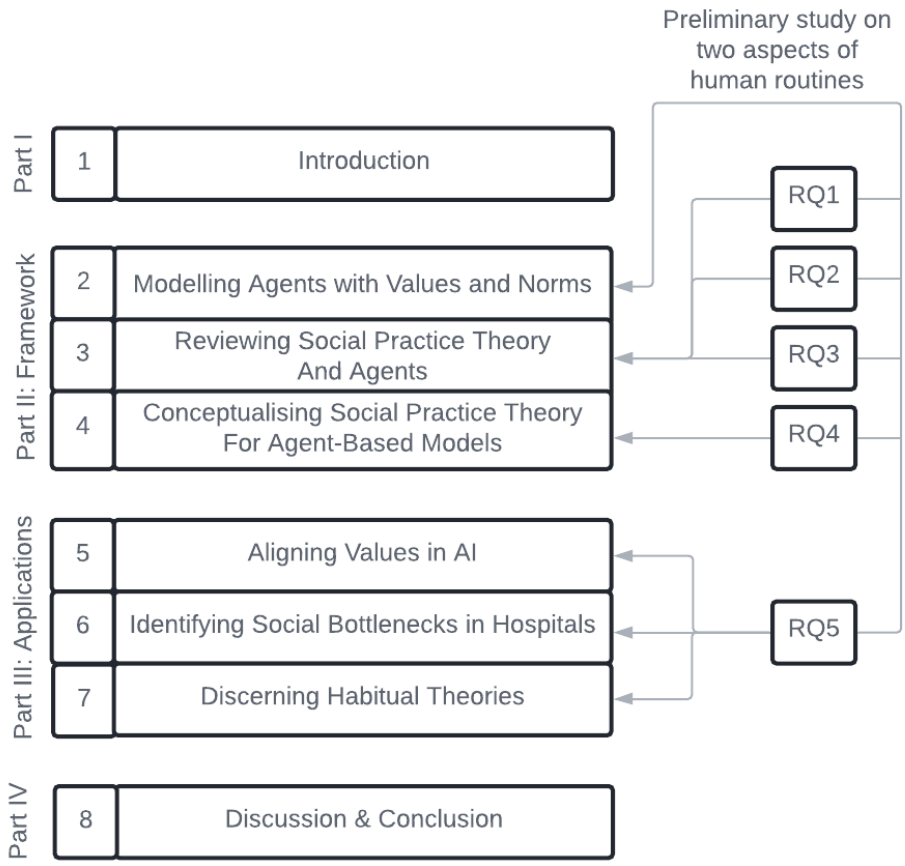


Figure 1.2: Thesis outline diagrams that shows four parts are split up into eight chapters and connected to the five research questions.

by using different parts of SoPrA.[3]

Chapter 5 presents a study that focuses on the value-alignment problem: a well-known problem in AI where one aims to match the behaviour of an autonomous agent with human values. The study analyses to what extent an autonomous agent is able to estimate the values and norms of a simulated human. Our framework supports this study by enabling the simulation of humans with values and norms.

Chapter 6 presents a study on identifying social bottlenecks in hospitals. The study provides a formal model (an OWL ontology) of the social dimension of an emergency room (ER). We use a formal reasoner to automatically identify whether certain social guidelines are being met. For example, the tool is able to identify for which activities the staff does not show an understanding of certain needs of the patient. Our framework supports this study by capturing key aspects of the ER in ABSS: agents follow protocols (routines) and make coordinated decisions deviating the protocol.

Chapter 7 presents a study comparing two theories on how habits break via simulation. Inducing behavioural change requires a good understanding of how habits break but multiple theories explain habituality. Via simulation, we identify an empirical experiment that enables social scientists to find out which theory is supported by evidence. Our framework supports this study by allowing the simulation of habitual behaviour.

Applying the framework allows us to reflect both on the application domain and the framework. The insights in the application domain show that the framework is applicable. The insights about the framework show the scope of the framework and the focus of future work.

Part IV discusses and concludes our thesis. First, we discuss the central choices made in our thesis. We provide arguments for these choices, relate them to choices made by other scientists or scientific fields, and in some cases advice future work to choose differently than we did in this body of work. Second, we discuss the scientific and societal relevance of our thesis. Last, we summarise the conclusion of our thesis in relation to our research questions.

This thesis ends with two appendices. Appendix 8.3 provides an overview of the requirements for the SoPra Framework, different computational versions of SoPrA, SoPrA extensions and the simulation models and their online location. Appendix 8.3 presents a preliminary study on mitigating gossip in organisations. The study provides a cognitive agent-based model to study the impact of various intervention strategies to control rumours in organisations. SoPrA extends current models on rumour-mongering by capturing how the behaviour of spreading rumours is habitual, social and interconnected with SPs as working and moving around. This appendix can be extended to a full fledged ABSS study by running the simulation experiment proposed in Section 8.3 and evaluating its results.

[3]The chapters use the at-that-time most recent version of SoPrA. These versions differ from the final version presented in Chapter 4. We reflect on this process in Chapter 8.

II

An Agent Framework for Human Routines

2

A First Step: Modelling Agents with Values and Norms

The sage puts himself last and becomes the first, Neglects himself and is preserved. Is it not because he is unselfish that he fulfills himself?

Laozi, Dao de Ching

This chapter has been published as R. Mercuur, V. Dignum and C. Jonker, "The value of values and norms in social simulation," J. Artif. Soc. Soc. Simul., vol. 22, no. 1, pp. 1–9, 2019 [174].

2.1. Introduction

Social simulations gain strength when explained in understandable terms. This paper proposes to explain agent behaviour in term of values and norms following [16, 60, 125]. Values are generally understood as 'what one finds important in life', for example, privacy, wealth or fairness [209]. Norms generally refer to what is standard, acceptable or permissible behaviour in a group or society [91]. Using values and norms in explanations has several advantages: they are shared among society [125], they have moral weight [209], they are applicable to multiple contexts [51, 182] and operationalized [229]. Moreover, humans use values and norms in folk explanations of their behaviour [161, 183]. Agents that use values and norms could thus lead to social simulation results that meet human needs for explanations.

To understand the relevance of agents with values and norm for social simulation, we need to know to what extent they can represent humans. Models are always simplifications from the system they are meant to represent, but understanding these differences clarifies the relevance of the model. Previous research primarily focussed on *constructing* agents that use human values and norms in their decision-making [16, 51, 60]. They gained insights in possibilities to synthesize theories on values and norms or how to formally argue in favour of an action in terms of values and norms. This paper aims to take the next step by comparing empirical data on human behaviour to simulated data on agents with values and norms.

We approach this by creating four agent models: a Homo economicus model, an agent model with values, an agent model with norms and an agent model with both values and norms. By comparing several agent models we gain more insight into the relative properties of the models. We do not expect the models to fully reproduce human behaviour, but we want to know how they compare and in what expects they differ. In particular, the Homo economicus model is used as a baseline as it is often used to represent humans (e.g., in game theory).

We simulate the behaviour of the agents in the ultimatum game (UG). In the UG, two players (human or agent) negotiate over a fixed amount of money ('the pie'). Player 1, the *proposer*, demands a portion of the pie, with the remainder offered to Player 2. Player 2, the *responder*, can choose to accept or reject this proposed split. If the responder chooses to 'accept', the proposed split is implemented. If the responder chooses to 'reject', both players get no money. We use two different scenarios: a single-round scenario and a multi-round scenario. In the single-round scenario, we test if the human behaviour can be reproduced by letting the agents evolve and converge to stable behaviour. In the multi-round scenario, we test if the change in behaviour humans display over multiple rounds of UG-play can be reproduced by the different agent models.

We compare the simulated data to empirical data from a meta-analysis that studied how humans play the ultimatum game. We focus on aggregated results: the mean and standard deviation of the demands and acceptance rate. We find that based on these measures a combination of agents with values and norms produces aggregate behaviour that falls within the 95% confidence interval wherein human plays lies more often than the other agent models. Furthermore, we find specific cases (responder behaviour in the multi-round scenario) for which agents

with values and norms cannot reproduce the learning nuances human display. We interpret this result as showing that agents with values and norms can provide understandable explanations that reproduce average human behaviour more accurately than other tested agent models. Furthermore, it shows that social simulation researchers should be aware that agents with values and norm can differ from human behaviour in nuanced learning dynamics. We find several insights on what aspects agents with values and norms outperform agents with solely values, the role of values and norms as static and dynamic components and how norms can produce different behaviour in different cases. We discuss the generalizability of these results given the dependence of these results on our translation from theory to model, parameter settings, evaluation measures and the use case.

The remainder of the paper is structured as follows. The next section presents theories on how agents use a Homo economicus view, values, norms or, values and norms in their decision making. Section 2.3 presents the two UG scenarios and the data on human behaviour in these scenarios. Section 2.4 presents our translation from theories to domain specific computational agent models. Section 2.5 presents the simulation experiments and the resulting behaviour of the different agent models. Section 2.6 discusses the interpretation and generalizability of these results.

2.2. Theoretical Framework

We use theories on the Homo economicus, values and norms to model the simulated agents. These theories are briefly summarized in this section.

2.2.1. Homo economicus

The *Homo economicus* (HE) agent is the canonical agent in game theory [190] and classical economics [139], that only cares about maximizing its direct own welfare, payoff or utility. As the agent only cares about its own direct welfare it will accept any positive offer in the UG. Humans in contrast reject offers as high as 40% of the pie [203].

One approach to explaining these findings is by extending the HE agent model to incorporate learning [93, 222]. The core of this explanation is that humans have learned through the feedback of repeated interaction to reject low offers to force the proposer into making higher offers. In this view, humans can be represented as learning Homo economicus agents for which, roughly said, fairness only exist as an instrument for wealth.

Our theory on the learning Homo economicus encompasses:

LHE.1 Humans only care about maximizing their own welfare.

LHE.2 Humans can learn that forgoing short-time welfare might lead to a higher long-term welfare.

2.2.2. Values

We view values as 'what a person finds important in life' [209] that function as 'guiding principles in behaviour'. In the remainder of this subsection, we will describe some of the work on values in psychology, sociology and philosophy focusing on how we can use values in the decision making of agents.

Schwartz developed several instruments (e.g. surveys) to measure values [229]. Based on these measurements Schwartz [229] can distinguish ten different basic values: self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence and universalism. (These basic values, in turn, represent a number of more specific values like wealth and fairness.) Schwartz shows that although humans differ in what values they find important there is a general pattern in how these values correlate. For example, people who give positive answers to survey questions on wealth are more likely to give negative answers to survey questions on fairness. These findings on intervalue comparison have been extensively empirically tested and shown to be consistent across 82 nations representing various age, cultural and religious groups [26, 56, 92, 229, 230].

Values have a weak, but general connection with actions [104, 182]. Miles [182] used data from the European Social Survey to show that values predict 15 different measured actions over six behavioural domains and in every country included in the study. Gifford [104, p. 545-546] reviews environmental psychology and concludes that the correlation between action and values is consistent, but weak, such that moderating and mediating variables are needed to predict actions from values. Following this research, we view values as abstract fixed points that actions over many context can be traced back to.

When making a decision between two actions there might be a conflict between two values. For example, when choosing to give away money or to keep it one might experience a conflict between the value of wealth and fairness. Poel and Royakkers [209, p.177-190] discusses different ways to resolve a value conflict: a 'multi-criteria analysis' or threshold comparison. In multi-criteria analysis, the different actions are weighted on the values and compared on a common measure; in threshold comparison an option is good as long as both values are promoted above a certain threshold. If one action upholds both thresholds, while the other one does not the former is chosen. If both option uphold both thresholds, threshold comparison does not specify what option to take. This paper uses multi-criteria analysis as this allows our agent to always make a concrete choice and therefore serve as a computational model for simulation.

Our theory on values thus encompasses:

V.1 There are ten different basic universal values (i.e., self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence and universalism.) that each represent a number of specific values (e.g., wealth and fairness) Schwartz [229]).

V.2 Humans are heterogeneous in the values they find important.

V.3 The importance one attributes to these values is correlated according to the findings of Schwartz [229]. For example, the values of wealth and fairness

are negatively correlated.

V.4 Values are (for the aim of this study) the direct and only cognitive determiner for actions.

V.5 When values are at conflict in a decision, humans use a multi-criteria analysis to resolve the conflict.

2.2.3. Norms

We follow Crawford and Ostrom [53] in that norms have four elements referred to as the 'ADIC'-elements: Attributes, Deontic, aIm and Condition.[1] The attribute element distinguishes to whom the statement applies. The deontic element describes a permission, obligation or prohibition. The aim describes the action of the relevant agent. The condition gives a scope of when the norm applies. One example in the context of the UG can be found in Table 2.1.

Table 2.1: A norm decomposed according to the ADIC-elements.

A	D	I	C
Proposers	should	demand 60% of the pie	when in a one-shot Ultimatum Game

What norms exists in a scenario? We say a norm exists when it influence the behaviour of an agent. We follow Fishbein and Azjen [91] in that norms influence behaviour either because of perceptions of what others expect or what others do.[2] Fishbein and Azjen [91] use the term 'perceived norm' (or: subjective norm) to make clear that it is a person's individual perception that influences behaviour and that perceptions may or may not reflect what most others actually do or expect. Thus a norm exists, for a particular person, when that person perceives other people do or expect it. To put it in terms of the ADIC syntax: a norm exists, for a particular person, if and only if the person (in Crawford and Ostrom [53]'s terminology Attribute) perceives others do or expect the aim given that the condition holds.

Empirical work shows that there is a correlation between norms and action. For example, a meta-analysis on the theory of planned behaviour shows an average R^2 of .34 between subjective norms and intentions. In other words, a linear model that takes measurements on the subjective norm as input can on average explain 34% of the variation of the measured intentions [14]. (Intentions, in turn, can explain about half of the variance in behaviour.). There are many different theories on how this relation between action and norm precisely works. For the purpose of this study, we aim to explore to what extent we can explain agent behaviour using only an understandable concept as norms.

Our theory on norms thus encompasses:

[1] Crawford and Ostrom [53] distinguishes norms from rules. Rules differ from norms in that they have a *unique* sanction when one does not abide them. In the UG, there are predominantly norms at play and not rules as players can differ in the sanctions they apply: reject the offer or accept but lower their esteem of the opponent.

[2] Note that the concept of norm of both Fishbein and Azjen [91] and Crawford and Ostrom [53] overlaps with what is often called a social norms (as opposed to e.g. a legal or moral norm).

N.1 A statement is a norm if and only if it has the following four elements: Attributes, Deontic, aIm and Condition.

N.2 A norm exists, for a particular person, when that persons perceives other people do or expect it.

N.3 The action a human does is the same as what they perceive as the norm.

2.2.4. Values and Norms

We follow Finlay and Trafimowm [89] in that some humans use values while others use norms. In a meta-analysis covering 30 different behaviours, they found that some humans are primarily driven by attitude (which strongly correlates with values) and some individuals are primarily driven by norms. We choose this theory for its simplicity and postpone more complex combinations of values and norms to future work.

Our theory on norms and values thus combines our theory on values (**V.1** - **V5**) and norms (**N.1** - **N3**) and adds:

VN.1 Some humans always act according to the norm and other humans always act according to their values.

Note that **V.4** and **N.3** in the case of the third theory only apply to a subset of the agents.

2.3. The Scenario

In this section, we describe how humans behave in two UG scenarios. We will use the simulations to check if our models, which we will describe in the next section, can reproduce this behaviour.

The UG has been the subject of many experimental studies since its first appearance in [114]. In this study, we use the meta-analysis by Cooper and Dutcher [48] as our main data source for human behaviour. We obtained the data of 5 of the 6 studies from the authors, namely: [221], [238], [12], [117], and Cooper et al. [50]. We obtain a total of 5950 demands and replies with on average the following specifics:

- An experiment has 32 players: 16 proposers and 16 responders.
- The pie size P is 1000.[3]
- A proposer can demand any $d \in D = [0, P]$
- A responder can choose a reply $z \in Z = \{accept, reject\}$
- The players are paired to a different player each round, but do not changes roles.

[3] For ease of presentation we chose 1000 with no monetary unit to the pie size. Although empirical work [203] shows that the effect of the pie size is relatively small, in further work we need to check the critically of this assumption.

- Players are anonymous to each other.

These studies can be separated on the amount of rounds the subjects play. One round comprises one demand for each proposer and one reply for each matched responder. We consider two scenarios: the one-round ultimatum game and the multi-round ultimatum game.

2.3.1. Scenario 1: One-Shot Ultimatum Game

The ultimatum game where players only play one round is called the one-shot ultimatum game. We subset the dataset on first-round games and depict what humans do in these rounds in Table 2.2.

Table 2.2: First-round human behaviour according to our adapted dataset. We display the estimated average demand (with its confidence interval (CI)) and acceptance rate.

datapoints	demand (μ)	with CI	demand (σ)	accept (μ)	with CI	accept (σ)
310	562	547-576	129	0.81	0.76-0.85	0.40

One popular explanation of why humans make these particular demands and accepts is that they have learned this in repeated interactions with other humans. When scholars talk about this type of learning, they mean an evolutionary sort of learning that takes place over long periods of time. Debove et al. [59] reviewed 36 theoretical models that all aim to explain first-round UG behaviour with such an evolutionary model. The idea behind these studies is that one simulates many rounds of behaviour in the ultimatum game and checks if this results in the demands human make in one-shot games. [4] In Section 2.5, we will check if our theories can explain the data in a similar way.

2.3.2. Scenario 2: Multi-Round Ultimatum Game

The original study of Cooper and Dutcher [48] focuses on how behaviour of responders evolves over 10 rounds. In Figure 2.1, we use the obtained data to represent two of their main findings.

In the left figure, we see that the share proposers demand slightly rises over time. In the right figure, we see that the responder's acceptance rate slightly falls and then rises. According to Cooper and Dutcher [48], the behaviour in the first five rounds significantly differs from the behaviour in the last five rounds.[5] Although the differences are small Cooper and Dutcher analyze them as they believe they can be informative. They assume that the mechanisms that are responsible for the change in behaviour over time, are also the mechanisms that bring about the behaviour in the first round. In Section 2.5, we present experiments that check if our theories can explain this change in behaviour over time.

[4] The catch here, is that these scholars do not believe that humans have played ultimatum games since the dawn of time, but that they have learned to make fair demands in (ultimatum-game-like) life experiences. Humans then display this behaviour at the first round of the actual psychological experiment. This is in contrast with the multi-round scenario were the simulation is actually compared to multiple rounds of real human ultimatum game play.

[5] Note that for a full statistical analysis we will need an ANOVA-test. For our purposes, it is enough to concern ourselves with the findings of [48].

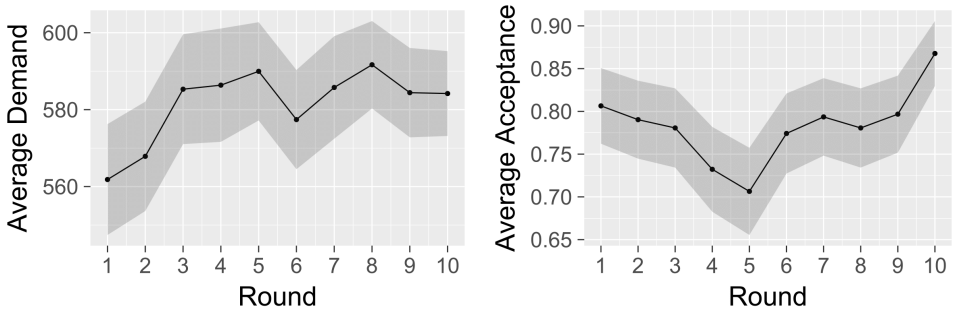


Figure 2.1: Multi-round human behaviour according to our adapted dataset. We display the estimated average demand (left) and acceptance rate (right) for different rounds. The grey area depicts the 95% confidence interval.

2.4. Model

If we want our results to be relevant for our theory (instead of an ad-hoc model), we need to be clear about the relation between the theory and a domain-specific model. In this section, we present our ultimatum-game specific implementation of our normative and value-based agent theory. The normative model has been implemented in Repast Java [198], the value-based model has been implemented both in Repast Java and in R for verification. The code, documentation and a standalone installer are provided at Github (see Appendix 8.3).

2.4.1. Learning Homo economicus Agent

In case of the learning Homo economicus agent there are already a few models available that can be applied to the UG. This paper uses the reinforcement learning models presented in [222] and [84], because in our view they focus on the core mechanisms of the Homo economicus and they are well documented.

In these models, each player keeps track of an utility u for a range of portions of the pie A (in our case $A = \{0, 0.1P, 0.2P, \ldots, P\}$.) For the proposer, this number represents the demand it makes. For the responder this number represents a threshold; if the demand is above this threshold it will reject, if the demand is equal or below the threshold it will accept. The model is initiated by letting each player (n) attribute an initial utility (i) to each pie-portion $a \in A$, such that, $u_n(t = 1, a) = i$.

Each round the agent does the following:

1. Each round a player picks a pie-portion according to the distribution of these utilities. In other words, the probability H, to pick a pie-portion a, is defined by the following function $H(a) = \frac{u_n(t,a)}{\sum_{a \in A} u_n(t,a)}$.

2. The proposer's demand is equal to that chosen pie-portion. The responder accepts the demand if its below its chosen pie-portion and rejects otherwise.

3. Each player n updates the utility u_n of the played action $\hat{a}$ by adding the obtained money r to the previous utility, i.e. $u_n(t + 1, \hat{a}) = u_n(t, \hat{a}) + r$. The

utility of the other actions remains the same.

Erev and Roth [84], Roth and Erev [222] present two versions of the Homo economicus that differ in their approach to the initial utilities. Before introducing them we first introduce the parameter $s(1)$, the initial strength of the model, defined as the ratio between sum of the initial utilities and the average reward, i.e. $s(1) = \frac{\sum_{a_i \in A_i} u(a)}{0.5P}$. The initial strength determines the initial learning speed of the agent. The two versions of the model are:

1. The initial utilities are all equal to each other i.e. u(a) = u(b) for all actions $a, b \in A$. (But $s(1)$ is free.)
2. The initial utilities sum to 1, i.e. s(1) = 500. (But are randomly distributed.)

For pragmatic reasons, we aim to test only one of the models to reproduce human behaviour. Neither of the models presented by Erev and Roth [83] or Roth and Erev [213] show the explicit reproduction of the UG results. In [84] the authors show that for many games data can be reproduced with the simple reinforcement learning agent introduced and equal initial utilities, but do not treat the UG in this paper. In [222], the authors show that crudely UG results can be reproduced with random utilities and a fixed strength of 500, but do not provide the exact parameter settings nor specifically compare the learned distributions to first round play. In this study, we choose to further explore the first model (with equal utilities) as the parameter space is more manageable. Future work should explore other reinforcement models including versions where one can vary the learning rate of the agents.

We now aim to specify which extra assumptions have been made when translating the theory to a domain-specific model:

LHE+.1 Players attach utilities to pie-portions that represent the demand for the proposer and a threshold for the responder.

LHE+.2 The initial utilities for these pie-portions are all equal to each other in the first round.

LHE+.3 There is a one-to-one relation to the utility of a pie-portion and the sum of the rewards it got you (e.g., no discount factor or utilities attached to sequences of actions).

2.4.2. Value-based Agent

Given V.1 there are ten basic values that each represent a number of specific values. In the context of the UG, we assume that the value of wealth and fairness are more relevant than other values. This is an educated guess based on that the behavioural economics literature frames the decision in these terms [49] and the meaning we associate with the values of wealth and fairness.

Given V.2 humans are heterogeneous in the values they find important. We represent this in the model by a parameter i_v that represents the importance (or weight) one attributes to the value.

Given V.3 this importance is correlated according to the findings of [229]. According to [229] the two values are strongly negative correlated. For pragmatic reasons, we will assume these values are perfectly negative correlated. This allows us to simplify the model to have two parameters μ and σ that specify a normal distribution from which the *difference* (di) in value strengths is drawn, i.e. for every agent

$$i_w = 1.0 + 0.5di \tag{2.1}$$

and

$$i_f = 1.0 - 0.5di \tag{2.2}$$

such that, $di = i_w - i_f$, represents how much more an agent values wealth over fairness.

Given V.4 values are the only cognitive determiner of actions. To make a computational model, we propose a procedure where the agent attributes a utility to every action and chooses the action with the highest utility. This utility should be determined by both the value of wealth and fairness. In other words, the agent will do a multi-criteria analysis to decide on the best action (V.5).

We present the decision-making model in three steps: (1) we relate to what extent a value is satisfied by the resulting money the agent obtains in one round of UG-play; (2) we relate this value-satisfaction and the importance one attributes to the value to a utility per result; (3) we relate this utility to the action the agent chooses.

First, to relate to what extent a value is satisfied by the resulting money of the agent, we have to interpret the meaning of wealth and fairness. Given the meaning of wealth, we assume that the higher one values wealth the higher the demands one makes (and expects). Given the meaning of fairness, we assume that the higher one values fairness the more equal the demands one makes (and expects). We represent this in the following function:

$$s_w(r) = \frac{r}{1000} \tag{2.3}$$

$$s_f(r) = 1 - \frac{|0.5P - r|}{0.5P} \tag{2.4}$$

where s_x specifies the extent to which the resulting money (of one round UG) r satisfies value x and P is the pie size. The satisfaction of wealth thus increases as one gets more money and the satisfaction fairness peaks around an equal split. [6]

Second, to relate this value-satisfaction (s) and value importance (i) to a utility (u) per result (r), we can combine s and i in several ways. This paper evaluates three possibilities. A divide function

$$u(r) = -\frac{i_w}{s_w(r) + ds} - \frac{i_f}{s_f(r) + ds}, \tag{2.5}$$

[6] Note that we chose to model the denominator as 1000 and not as P; the rationale is that we think the satisfaction of wealth increases absolutely and not relative to the pie size. In further work, we should further explore empirical work to support this modelling choice.

a product function

$$u(r) = i_w * s_w(r) + i_f * s_f(r), \tag{2.6}$$

and a difference function

$$u(r) = i_w - s_w(r) + i_f - s_f(r). \tag{2.7}$$

Every utility-function thus represents a different model. In the next section, we will evaluate which model can best reproduce human behaviour.

Third, to relate this utility to the action the agent chooses we postulate that:

- the proposer now demands that $d \in [0, P]$ for which the utility (as given by $u(r)$) is maximal.
- the responder chooses to accept if (and only if) the utility of what it receives - $u(P - d)$ - is higher than the utility of a reject.

We choose to model the utility of rejection by filling in the chosen utility function with $s_w(0)$ and $s_f(0.5P)$, i.e. the agent interprets it as getting maximum fairness (as in the r= 0.5P case), but getting almost no wealth (as in the $r = 0$ case).

In summary, to translate our theory to our domain we have added the following parts to our theory:

V+.1 Wealth and fairness are the only relevant values in the ultimatum game.

V+.2 The importance one attributes to wealth and the importance one attributes to fairness are perfectly negatively correlated.

V+.3 The higher one values wealth the higher the demands one makes (and expects). The higher one values fairness the more equal the demands one makes (and expects).

V+.4 Humans compare to what extent wealth and fairness are satisfied by

(a) a divide function (function (2.5)).

(b) a product function (function (2.6)).

(c) a difference function (function (2.7)).

2.4.3. Normative Agent

Given **N.1** a statement is a norm if and only if it has the attribute, deontic, aim and condition element. In the context of the ultimatum game we consider the types of norms stated in Table 2.3. Note that according to our theory, all sorts of possible norms could be considered. For example, 'responders should reject in all cases'. We consider only the type of norms in Table 2.3 as we think those are likely to exist in this domain.

To know which norms actually exist in a particular game, we look at **N.2**. This part of our theory states that a norm exist, for a particular person, when they perceive other people do or expect it. Note that in our scenario an agent does not

Table 2.3: The norms considered in the ultimatum game split out according to the ADIC-syntax, where p refers to a proposer, q to a responder, d to a demand and t to a threshold.

label	A	D	I	C
$N_{p\hat{d}}$	Proposers	should	demand $\hat{d}$	in the UG
N_{qt}	Responders	should	reject	if and only if the demand is above threshold t in the UG

switch roles (i.e. proposers stay proposers, responders stay responders). Proposers thus never see the actions other proposers do, but can only rely on what they think responders expect (from proposers). The situation is analogous for responders. The question is thus, how does one derive what the opponent expects from you given his or her actions?

In the case of the responder this is fairly straightforward. What does a proposer expect from a responder when demanding X% of the pie? He or she probably expects that the responder would accept that demand (and lower), but reject everything higher than that. In other words, the demand becomes a certain threshold for acceptance. For multiple rounds, we assume this threshold is calculated by averaging over all seen demands. Formally, this amounts to that norm N_{qt} exists for responder $q \in A$ and a threshold $t \in D$ if and only if

$$t = \frac{\sum_{d \in OD} d}{|OD|}, \tag{2.8}$$

where OD are the demands responder r has observed in the games it participated.

For the proposer, it's a bit more tricky to deduce what behaviour is expected. We postulate that the demand a proposer is expected to make is equal to the average of two indicators: the lowest demand that is rejected and the highest demand that is accepted. Formally, this amounts to that the norm N_{pd} exists for a proposer $p \in A$ and demand $\hat{d} \in D$ if and only if

$$\hat{d} = \frac{\min_{d \in RD} d + \max_{d \in AD} d}{2}, \tag{2.9}$$

where RD is the set of demands that the proposer p has seen rejected and AD is the set of demands that the proposer p has seen accepted.

For most cases the action of the proposer and responder is now clear: they act according to what they perceive as the norm (**N.3**). However, our theory does not specify what agents should do when they perceive no norm. For the sake of making a computational model we postulate that if no norms exist the agent draws a random action from a uniform distribution. Section 2.5 explores uniform distribution with different means to gain insight into the relevance of this assumption on our results.

We postulate that if no norm exist the agent does a random action. Note that to translate our theory to our domain we have added the following parts to our theory:

N+.1 Proposers expect that responders accept their demands, but reject everything higher than that.

N+.2 Responders expect that proposers demand the average of the lowest demand that is rejected and the highest demand that is accepted.

N+.3 If no norms exist, then humans draw a random action from a uniform distribution.

2.5. Experiments & Results

In this section, we test four agent models in both the one-shot and multi-round scenario. We evaluate the models on their ability to reproduce human behaviour by comparing the 95% CI wherein human behaviour lies to the simulated behaviour.

2.5.1. Reproducing First-Round behaviour

To test our theories on their ability to reproduce human first-round behaviour we let the agents interact until their behaviour stabilizes. To set-up this experiment we thus need to simulate a number of 'pre-rounds'. We assume that these 'pre-rounds' are similar to the scenario as described above. For example, the amount of players is 32 and the agents do not switch roles. If the stable behaviour is the same as the human first-round behaviour, then the theory serves as an explanation of how humans have learned to make the demands and rejects they display.

We find that if we average over 100 runs per parameter set-up the confidence interval around the estimated means is very small. The remainder of this section thus treats the estimated mean as the true mean.

Testing our learning Homo economicus model

We test our learning Homo economicus model on its ability to reproduce first round behaviour in the UG. We can run different versions of the model dependent on the initial utilities the agents attribute to their actions (which are given by parameter s). Using exploratory simulations we find that behaviour stabilizes around pre-round '500'.

We run simulations for $s \in [0.00005, 8]$ with a logarithmic stepsize as exploration learns that the result of simulations outside this interval do not significantly differ from the result of the bounds. We calculate for each parameter set-up the distance between human demand and acceptance rate and the simulated demand and acceptance rate and find that for $s = 0.03$ this distance is minimal; the results for this parameter setting are displayed in Table 2.4. Furthermore there is a negative exponential relation between s and the distance.

Table 2.4: The demand and acceptance rate of the lhe agent. Note that 'avg.' and 'sd.' refer to the average and standard deviation of the demand and acceptance rate over *one run*.

avg. demand	sd. demand	avg. accept	sd. accept
551.2	258.7	0.60	0.10

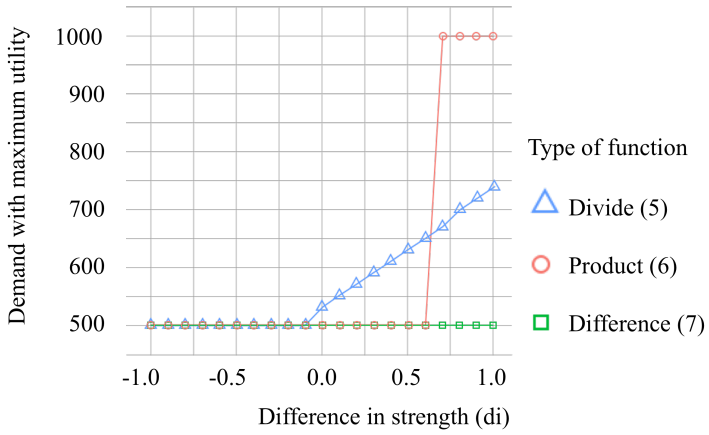


Figure 2.2: Three value functions compared on what demand gives maximum utility (y-axis) for different value strengths (x-axis).

We conclude from Table 2.4 that the learning Homo economicus agent particularly differs from humans in the distribution of demands and acceptance rate. Although the learning Homo economicus can reproduce human demands it can only do this when other agents force it into making lower demands by rejecting enough. This model cannot explain why proposers make relatively equal demands while responders accept almost all demands (81%).

Testing our value-based agent model

We test our value-based agent model (**V**) on its ability to reproduce first-round behaviour in the UG. We can run different versions of our value-based agent depending on which function the agent uses to combine the satisfaction of different values (**V+.4**) and with which μ and σ the difference in value strength (di) is normally distributed.

By calculating which value-based agent model leads to which distribution of demand we can gain insight in which model best reproduces the demands human make. Figure 2.2 compares the three different agent models by showing what the best demand for each agent is given the importance it attributes to its values. Recall, that human demands are normally distributed. Given that di is normally distributed, we can see that the divide function is the only function for which the demands agents make will be normally distributed. We conclude that our value-based agent model with extension **V+.4**a has the best chance of reproducing human behaviour.

To find out if the value-based agent can reproduce the demand and acceptance rates human display we simulate the agent.[7] We run experiments for $\mu \in [-2, 2]$ and $\sigma \in [0, 2]$ with stepsize 0.01 and denote the average demand and the aver-

[7] Note that in the case of the value-based agent the behaviour stabilizes in round 1 as the agents do not learn.

age reject rate they result in. We calculate for each parameter set-up the distance between human demand and acceptance rate and the simulated demand and acceptance rate (i.e., the error). We find that for $\mu = -0.55$ and $\sigma = 1.14$ the distance is minimal; the results for this parameter setting are displayed in Table 2.5. Note that the resulting behaviour fall within the 95% CI around whic the human play lies (see Table 2.2). The distance between human play and simulated play increases with a linear relation to how far μ and σ move away from this optimal setting.

Table 2.5: The demand and acceptance rate of the value-based agents. Note that 'avg.' and 'sd.' refer to the average and standard deviation of the demand and acceptance rate over *one run*.

avg. demand	sd. demand	avg. accept	sd. accept
560.9	103.7	0.82	0.37

We conclude that our value-based model can for a specific parameter range quite accurately reproduce human demands and acceptance rates.

Testing our normative agent model

We test our normative agent model (**N**) on its ability to reproduce first-round behaviour in the UG. We can run different versions of our normative agent depending on what the agent does when no norm is specified (i.e. round 1). Using exploratory simulations we find that behaviour stabilizes around pre-round '15'.

In our first experiment, the normative agents draw their demand from $U(0, P)$ and their acceptance rate from $U(0, 1)$. Table 2.6 presents the demand and acceptance rate the agents demonstrate when their behaviour stabilizes. The average demand and acceptance rate clearly significantly differ from human play (see Table 2.2). The agents demand just a bit less than half of the pie, where the humans demand more than half. The agents have an accept rate of 0.5 and humans 0.85. This experiment gives some evidence that a normative theory cannot serve as an explanation for first-round behaviour. However, the stable behaviour is close to the initial behaviour: the mean of the uniform distributions the agents draw their initial actions from is close to 494.0 and 0.50. This raises the question how dependent our results are on the initial conditions (N+3) and if other initial conditions might reproduce first-round human behaviour.

Table 2.6: The demand and acceptance rate normative agents display when their behaviour is stabilized (pre-round 15). Note that 'avg.' and 'sd.' refer to the average and standard deviation of the demand and acceptance rate over *one run*.

avg. demand	sd. demand	avg. accept	sd. accept
494.0	178	0.50	0.5

To test if our normative model could reproduce human behaviour under different conditions, we run analogues experiments for different uniform distributions. Figure 2.3 depicts the resulting demands and acceptance rate for different initial demands and acceptance rates. We find that the resulting demands are fairly close to the initial demands. The demands can converge to higher or low than the initial demands depending on the initial acceptance rates. For some initial conditions

human demands can be reproduced. In contrast, the acceptance rate converges to either 0.5 or 1.0, but never comes close to the human 0.85. Although hard to display in this figure, inspection of the data shows that its the edge cases (e.g., where the initial demand is 0 or 1000) that convert to an acceptance rate of 1.0.

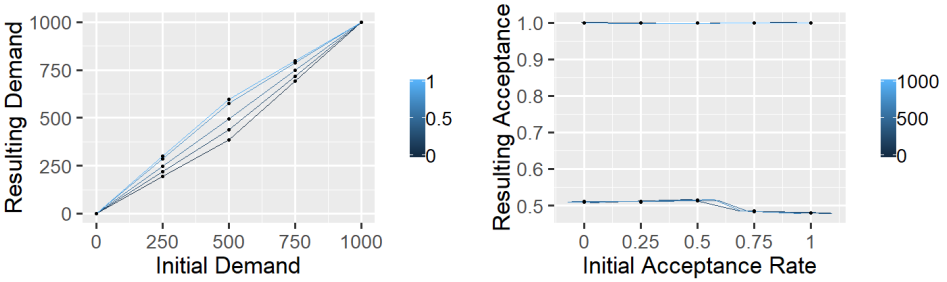


Figure 2.3: Left: the average resulting demand for different initial demands (x-axis) and different initial acceptance rates (color). Right: the average resulting acceptance rate (y-axis) for different initial acceptance rates (x-axis) and different initial demands (color).

To gain more insight in the role norms can have in human decision-making we highlight a few more aspects of these results. First, Figure 2.3 shows that although one might expect a normative agent model to 'normalize' both the demand and acceptance rate can converge on different values than they started. Second, table 2.6 shows a fairly large standard deviation for both the resulting demand and acceptance rate. This shows that although agents act according to a norm there are still individual differences per agent.

We conclude that our normative model can reproduce human demands, but not simultaneously reproduce human acceptance rates. The simulation shows though that normative models can have counter-intuitive results where resulting norms drift away from the original norm and where agents can have individual norms and reproduce a similar variance behaviour as humans do.

Testing a combination of normative and value-based agents

In our second experiment, we test if we can reproduce human behaviour with our theory that some people act according to their values, while others act according to their norms (**VN**). In Figure 2.4, we depict the demand and acceptance rate for different numberof normative agent.

We compare the average demand and acceptance rate of our agents against the confidence interval wherein human play lies (grey area).[8] As we already knew, we can reproduce human behaviour by simulating only value-based agents (under the specific parameters mentioned in 2.5.1). We find that we can also reproduce the demands if we allow up to half of the agents to act out of norms. This can be explained by that if we allow enough value-based agents to make realistic demands, the normative agents will learn to adhere to the norm these agents set. In contrast,

[8] To exactly conclude what amount of normative agents reproduce human behaviour we should do more rigorous statistical analysis (e.g. an ANOVA-test). However, for our purposes it suffices to look at the 95% confindence interval.

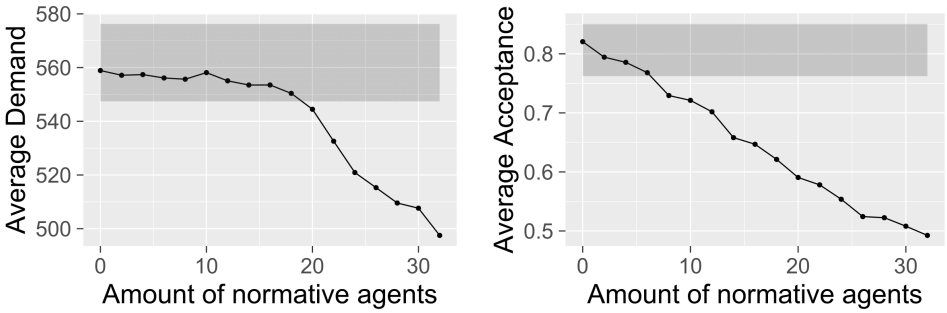


Figure 2.4: The average demand (left) and average acceptance rate (right) over all agents (y-axis) for different amounts of normative agents (x-axis). Note that if there are, for example, 10 normative agents, there will be 22 value-based agents.

only for a very small range of normative agents we can reproduce the acceptance rate as well.

We conclude that our theory on values *and* norms (**VN**) reproduces human demands and acceptance rate as long as the amount of normative agents is limited.

2.5.2. Reproducing Multi-Round behaviour

For multi-round behaviour, we are interested in the behaviour agents display in round 2-10. In contrast to the one-shot scenario, we have empirical data on what the real demand and acceptance rate is humans initially display (round 1). Therefore, we initiate the model based on the empirical data and then test if the agents reproduce human behaviour in subsequent rounds. We evaluate the simulated data on if it falls in the 95% confidence interval in which human play lies. In addition, we highlight aspect of the learning dynamics the agents display.

Testing our learning Homo economicus agent model

The learning Homo economicus agent is tested on its ability to reproduce multi-round behaviour in the UG. We assume here that the learning Homo economicus agent already reproduces human play in the first-round, and see if it can reproduce the learning process humans display. We can run different version of the model depending on the initial utilities the agent attributes to their action (which are given by parameter s)

We run simulations for $s \in [0, 50]$ as exploration taught us that the result of simulations outside this interval do not significantly differ from the result of the bounds. We depict the results in Figure 2.5. We can see that for both the average demand as well as the average acceptance rate the simulated behaviour does not fall into the confidence interval in which human behaviour lies. However, we can see some similarities in the learning dynamics between the simulated behaviour and the empirical data. In case of the proposer, on round 3, 5, 6 and 8, the simulated data shows a similar rise and fall as the empirical data for most values of s. In case of the responder, from round 5 onwards the simulated data shows a similar rise and fall as the empirical data for some values of s.

One explanation of these results is that the learning Homo economicus differs from humans in wanting to explore other (on average different) options than first-round behaviour. The other behaviour yields enough utility to not change it mind back.

We conclude that the learning Homo economicus agent cannot reproduce the average demand and acceptance rate humans display. For some values of s the learning Homo economicus agent reproduces some of the learning dynamic humans display.

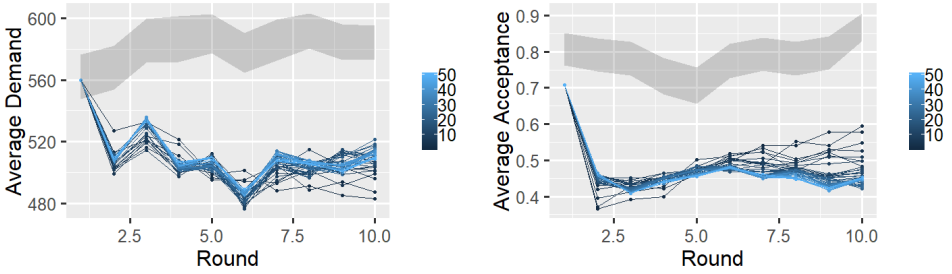


Figure 2.5: The average demand (left) and average acceptance (right) at different rounds. The colored lines depict the behaviour of the agents, where the color signifies the value of the s parameter (which influences the initial conditions). The grey area represents the 95% confidence interval wherein human play lies.

Testing our value-based agent

For our value-based agent model we can analytically see that it will not be able to reproduce the human dynamics of multi-round behaviour. The value-base agent behaviour does not change over time and it does not learn.

Testing our normative agent model

The theory on norms (**N**) is tested on its ability to reproduce multi-round behaviour in the UG. We assume here that the normative agent already reproduces human play in the first-round, and see if it can reproduce the learning process humans display. In other words, we adapt our normative agent such that it does not act randomly the first round, but does the demand and average accept humans do (i.e. we change **N+.3**).

In Figure 2.6, we depict the demand and acceptance rate of the normative agent over multiple rounds. We found that the average demands normative agents display is very similar to the average demands humans display. For almost all the individual points we can say with 95% confidence that they are the same as human play.[9] In contrast, the acceptance rates of the normative agents does not match that of humans. In the case of the proposer, the dynamics of the simulated data and the empirical data match in the general rise, but not in the small fluctations. In case

[9]This is not the same as being 95% confident that the two processes are the same. One way to check the similarity of time series is by fitting ARIMA-models to the two lines and compare those. However, for our purposes it suffices to look at the 95% confindence interval.

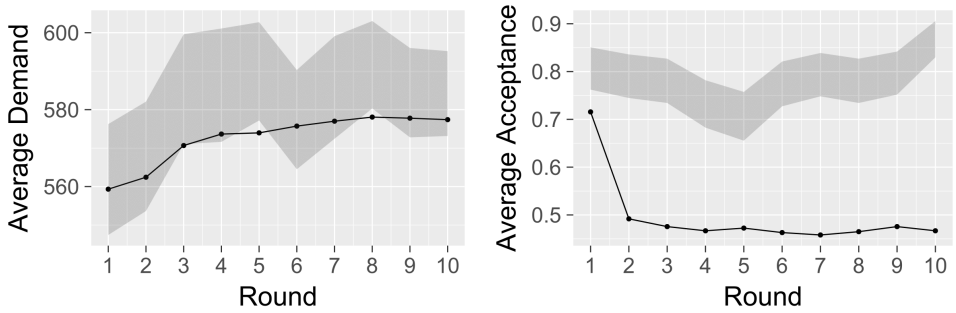


Figure 2.6: The average demand (left) and average acceptance (right) at different rounds. The black line represents the behaviour of the normative agents, the grey area represents the 95% confidence interval wherein human play lies.

of the responder, the dynamics primarily differ. In particular, the initial drop in the simulated data and the drop on round 5 of the empirical data have no counterpart.

We can explain the initial drop in acceptance rates as follows. After the first turn, the responders set their threshold to the first demand they saw (**N+.1**). After this, about half of the demands are above and about half of the demands are below this threshold leading to the 0.5 acceptance rate.

We conclude that the normative model cannot reproduce both human demands and acceptance rate. The learning of the proposer agent is similar to that of humans, but (at the same time) responder learning strongly differs.

Testing a combination of value-based and normative agents

In our last experiment, we test if we can reproduce human behaviour with our theory that combines both values and norms (**VN**). We depict the results in Figure 2.7. We found that no single combination of value-based and normative agent com-

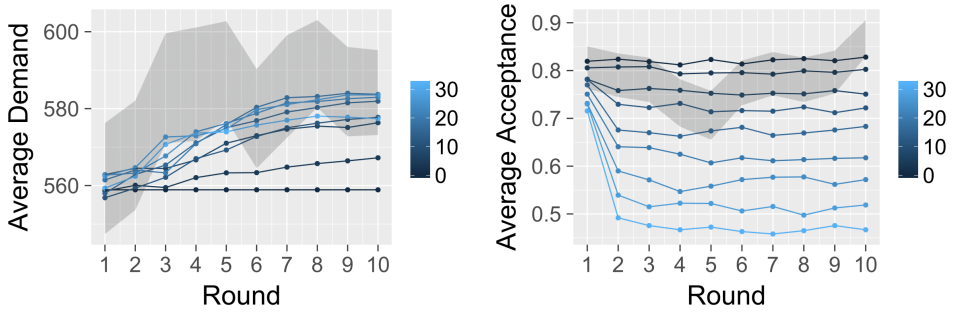


Figure 2.7: The average demand (left) and average acceptance (right) at different rounds. The colored lines depict the behaviour of the agents, where the color signifies the number of normative agents. The grey area represents the 95% confidence interval wherein human play lies.

pletely reproduces human play. However, when the amount of normative agents is

limited (e.g., 10) the proposer and responder learning is similar to that of humans. In this case, most of the simulated data points fall within the 95% confidence interval wherein human play lies. Note that in the proponent case human behaviour is best matched with a large majority of normative agents, while in the respondent case a majority of value-based agents gives the best fit. We find that the dynamics of the proposer agents, match the slight raise in demand humans display. The dynamics of the simulated responders differ from those of humans. In particular, a rise in acceptance rate after round 5 is not reproduced.

We conclude that no single combination of normative and value-based agent can completely reproduce human play, but that for some particular combinations the average demand and acceptance rate often lies within the 95% confidence interval humans display. In addition, some of the dynamics of the proposer or reproduced. The dynamics of the simulated responder strongly differs from human play.

2.6. Discussion

This paper compared empirical data on human behaviour to simulated data on agents to gain insight in to what extent agents with values and norms can represent human behaviour. We found that agents with values and norms can both evolutionary reproduce one-shot UG behaviour and, for most rounds, reproduces aggregate human behaviour in a multi-round scenario. Given the methodology of this paper, an agent with values and norms thus outperforms a learning Homo economicus agent or an agent that uses solely values and norms in its ability to reproduce human behaviour. It outperforms the learning Homo economicus agent and the normative agent in the one-shot UG and the learning Homo economicus, value-based and normative agent in the multi-round UG. The remainder of this section discusses to what extent this means agent with values and norms could represent human behaviour in explainable terms.

The generalizibility of these results is namely dependent on the fact that they hold under a specific translation from theory to model, specific evaluation, specific parameter settings and a specific use case.

We aimed to be transparent in our translation from theory to model so that other researchers can pinpoint on what aspects they agree and disagree and how these aspects influence the results. On aspects of the model we found necessary out of a pragmatic viewpoint, but not fundamental to the theory we aimed to study their influence on the output. For example, we showed that the normative model cannot reproduce human acceptance rate independently of its initial conditions. We hope this paper can serve a discussion between the social and computational sciences to pinpoint essential and accidental properties of values and norms.

We evaluated the agent models on to what extent aggregate measures of simulated behaviour fall inside a 95% confidence interval wherein aggregate measures of human behaviour lie. We find that based on this evaluation measure, we can differentiate models that cannot possibly reproduce human behaviour from models that can. For example, our learning Homo economicus agent cannot (under any parameter setting) come close to both humand demand and acceptance rate in the one-shot UG. Our value-based agent and agent with values and norms can repro-

duce human demand and acceptance rates. However, this latter result depends on specific parameter settings for these models (i.e., difference in values follow a certain normal distribution and there are a specific amount of normative agents). In future work, these parameter settings can be evaluated by using empirical data on how values are normally distributed and how many humans act out of norms (i.e., what Moss and Edmonds [188] call cross-validation). Based on this, we argue that interviews that measure how humans individually explain results would be a valuable addition to the data available on the UG.

We compared our simulated agents on data obtained on human behaviour in the UG: a lab experiment. Lab experiments have the advantage of allowing measurements in a reproducible controlled settings, which is the reason we were able to obtain a relatively large homogeneous dataset of a meta-study. Proponents of natural decision making criticize the lab setting as being unrepresentative for real-life decision making [147]. In the case of the UG, it is indeed unclear what real-life process the ultimatum game is exactly meant to represent. This has at least two consequences. First, it is unclear what modelling assumptions should be made about the setting (e.g., what do the initial conditions means in an evolutionary setting). Second, in a setting with unstable conditions, high stakes and more uncertainty humans could use a different decision-making process than in the UG. Future work should balance findings from this lab context with findings in a more natural context.

By comparing an agent with values and norms to other agent models we gained several insights. First, we found that in the one-shot UG a value-based agent can reproduce human behaviour just as well as an agent model with values and norms. We conclude that both agent models can be used to explain aggregate human behaviour in this scenario. Note that although an agent model with values and norms introduces another concept, a reason to choose it over a value-based model is that it can explain behaviour in a more general context (i.e., the multi-round scenario) or fits better with explanations humans give (e.g., in interviews). Second, the experiments show that values act as a static component that anchors the agent to certain behaviour, where norms form a dynamic component that can allow agent behaviour to drift away from human behaviour (e.g., in the one-shot UG) or towards human behaviour (e.g., in the multi-round UG). This gives insight into the role values and norms can have in humans and agents (i.e., as anchors and as dynamic learning components). Furthermore, this is relevant for creating ethical AI: AI that we create according to our ideas of values and norms can still drift away from behaviour we find acceptable. Third, we found that normative agents can still individually differ in behaviour just like humans. This is because agents can form different ideas of 'the norm' based on the different interactions they had.

Although social simulation focuses on reproducing general aggregate patterns in human behaviour, we should be aware that there are aspects of human behaviour agents with values and norms do not reproduce. The most notable difference is that nuances in the learning dynamics of the proposer and (especially) the responder behaviour in the multi-round scenario are not reproduced. Furthermore, if one is primarily interested in nuanced learning dynamics, then this research suggests that

other agent models should be used over agents with values and norms. For social simulation researchers, this means that they should be aware of possible differences between their agents with values and norms and humans in learning dynamics.

We suggest a few directions for future work to find agent models that on an aggregate level reproduce human behaviour. First, we could specify in more detail when to use norms and when values.[10] The results in the multi-round scenario show that for the proponent case human behaviour is best matched with a large majority of normative agents, while in the respondent case a majority of value-based agents gives the best fit. We are careful to not conflate this with the claim that proposers mainly use norms while responders mainly use values. The average demand, acceptance rate and the different rounds all depend on each other. There are simulation runs with predominantly normative agents (that explain proposer behaviour) and value-based agents (that explain responder behaviour), but this is not the same as one run where both results are explained simultaneously by agents first proposing out of norms and then the same agents responding out of values. It could very well be that the low acceptance rate in runs with a high amount of normative agents is due to the fact that in that same run the demands are higher (and not because the agents use norms). Future work should use simulations to check if agents that sometimes use values and sometimes use norms can explain aggregate UG behaviour.

Second, there are other deontic operators than 'should' that can be used to improve the normative agent model [271]. For example, the deontic operator 'may' can represent a range of possible demands the proposer considers permissible. Future work, could test to what extent these different deontic operators allow the normative agent to reproduce human aggregate behaviour.

Third, in experimental economics, there are several models that have been used to reproduce human behaviour [143]. As discussed, one of these models, the learning Homo economicus outperforms the agent with values and norms by being able to reproduce some of the nuanced learning dynamics. Future work, could look at how to combine the benefits of both models to reproduce human behaviour more accurately.

Fourth, Henrich et al. [122] explores the relation between 15 small scale societies and the actions taken in the UG. They find that market integration and economic differences explains a substantial portion of the behavioural variance across societies. This raises the question whether the variance in values (related to market integration) between cultures provide further explanation of the found behavioural variance. By extending the agent models provided here with a concept of culture, we enable ABM future research on the relation between culture, values and behaviour in the UG.

Last, considering that the motivation for this work is to use understandable terminology to explain the agent behaviour, we are more inclined towards using concepts humans use in their explanation. For example, Fehr and Fischbacher [87]

[10] One advantage of agent-based models is that we do not have to restrict our theories to some linear combination of values and norms (as much of psychology does), but can theorize any functional connection between them [45].

suggested that the concept of reputation can explain the learning dynamics in the UG: a responder could aim to build a reputation as a strong ('selfish') player.

2.7. Conclusion

This paper aimed to compare empirical data on human behaviour to simulated data on agents with values and norms. We found that agents with values and norms can both evolutionary reproduce average one-shot UG behaviour and, in most rounds, reproduces the average demands and acceptance rates humans display in a multi-round scenario. We interpret this result as showing that agents with values and norms can provide understandable explanations that reproduce average human behaviour more accurately than other tested agent models (e.g., the Homo economicus).

We gained several insights into the role of values and norms in agent models. First, we found that our agents with values and norms cannot reproduce the nuanced learning dynamics humans display (in particular the responder behaviour in the multi-round scenario). Second, we found that agent models with solely values or solely norms can reproduce some human behaviour in one scenario, but to reproduce behaviour in both scenarios a combined model is necessary. Third, the experiments show that values act as a static component that anchors the agent to certain behaviour, where norms form a dynamic component that can allow agent behaviour to drift away from human behaviour (e.g., in the one-shot UG) or towards human behaviour (e.g., in the multi-round UG). Fourth, normative agents can still individually differ in behaviour just like humans, because agents can form different ideas of 'the norm' based on the different interactions they had.

We discussed the dependence of these results on our translation from theory to model, parameter settings, evaluation and use case. Future work, should be directed at pinpointing essential and accidental properties of values and norms, interviews that measure how humans individually explain results to validate micro aspects of the model and balance findings from this artificial lab context with findings on natural decision-making.

Our study is a first step that shows how agents with values and norms can provide an improvement over simpler models in representing human behaviour in explainable terms.

3

Reviewing Social Practice Theory and Agents

This chapter has been published as R. Mercuur, V. Dignum, and C. Jonker, "Integrating Social Practice Theory in Agent-Based Models: A Review of Theories and Agents," IEEE Trans. Comput. Soc. Syst., 2020 [176].

3.1. Introduction

Evidence-driven agent-based modelling plays a useful part in understanding social phenomena [29, 80, 81, 187, 243]. This includes bearing evidence on human-decision making on our models for human decision-making: agents. To utilize the evidence of sociological and psychological research, Jager [134] argues for integrating socio-cognitive theories in our agent models. One of these socio-cognitve theories, called social practice theory (SPT), fits agent-based modeling as both study the interaction of humans with their social environment and the direct and indirect effect of these interactions on society. By grounding agent models in socio-cognitive theories, and in particular on SPT, we stand on the shoulders of years of sociological and psychological research on modelling human decision-making.

SPT provides a theory that describes our 'everyday doings and sayings' [227] and emphasizes that these so called social practices (SPs) are habitual, social and interconnected [34, 214, 227, 235]. Our day is full of SPs: working, dining, commuting, teaching, meeting, walking or sports. First, SPs emphasize that behaviour is habitual[34]. For example, when one is at the office, one habituality enacts the SP of working. Habituality helps us to understand why it might not be so easy to fall into the same working practice at home, how you intentionally go to the office to trigger the practice of working or how you can (with willpower) also develop a habit to work at home. Second, SPs emphasize that behaviour is social [214, 227]: a practice is not only individual but others have a similar practice. For example, when your colleague enters your office, he or she does not distract you but waits until the coffee break at 10.30 to discuss current matters. Sociality helps us to understand how your colleague concludes to wait until the coffee break based on her own practice: she believes reading is a kind of work, work promotes productivity, the coffee break starts at 10.30 and that you believe the same. Third, SPs emphasize that behaviour interconnected Shove et al. [235]. For example, your work-commute is connected to your sport-commute and you decide to take the car so you can do both[42]. Interconnectivity helps us to understand how you want both your work-commute and sport-commute to promote efficiency and therefore take the car or how your colleague understands that reading is connected to the practice of work and therefore promotes productivity. In short, SPs describe our everyday decisions and help us model the habitual, social and interconnected aspects of these decisions.

Dignum et al. [72], Kaminka [145] call for translating socio-cognitive theories to a domain-independent agent model to prevent researchers from reinventing the wheel. They identify that current agent-based models (ABMs) use similar socio-cognitive theories, but without a general framework researchers cannot reuse, compare or recombine these. This hampers both the efficiency [71, 145] and the evolution of models [80]. By translating SPT to a domain-independent agent model, we would enhance the comparability and reusability of agent models that model habituality, sociality and interconnectivity. However, for this purpose, we first need to identify what is required of an agent model that integrates SPs and if there is a gap in the current literature on agent models given these requirements.

This chapter provides a set of requirements for an agent model that integrates

SPT and verifies whether current domain-independent agent models satisfy these requirements. Previous work has emphasized the importance of SPT for agents [77], has made abstract models of SPT [127] or used SPT to study energy systems [191]. So far a review that lays out specific requirements for integrating SPT with agent models and evaluates current agent models does not yet exist. Such a review is useful for ABM researchers who are interested in integrating SPT with ABM and want to know in more detail what aspects SPT comprises, and what requirements these imply for implementations. Furthermore, it allows ABM researchers to pick one of the current agent models depending on their needs regarding habituality, sociality and interconnectivity. For this purpose, we distilled requirements from the literature on SPT, agent theory and social psychology. Each aspect (habituality, sociality and interconnectivity) is studied from the two perspectives that ABM aims to integrate: the individual agent perspective and the collective system perspective. We evaluated 11 agent models against the requirements we elicited. The selection of agent models is based on the review by [19], to which we added two more recent agent models. We split up the models in three categories: reasoning models (PRS, BDI, eBDI), normative models (BOID, BRIDGE, EMIL-A, NOA, MAIA) and social-psychological models (Consumat, PECS, Agent-0). By distilling requirements and evaluating the three categories of models, we provide an overview of the aspects SPT comprises and the current state-of-affairs in integrating these aspects in ABM.

We find that habituality, sociality and interconnectivity are empirical and theoretically grounded aspects of behaviour but have not been fully captured by current agent models. First, behaviour is habitual and often not conscious, voluntary or intentional [124, 186, 250]. We find that habituality is reflected in reasoning models by modelling reactivity and in the Consumat model by the ability of agents to repeat past behaviour. However, current agent models do not support (1) explicit reasoning about habits, (2) context-dependent habits and (3) individual learning concerning habits. Second, behaviour is intrinsically social and not individual with a layer of sociality on top of it [44, 71, 249, 263]. We find that sociality is somewhat reflected in models that use norms or that use social mechanisms (e.g., imitation), but current agent models do not use a comprehensive set of collective concepts, nor do they order social information around actions or relate individual and collective concepts (e.g., habits to norms) in order to guide interactions. For example, a normative agent model supports reasoning about the fact that most people work, but not that because an individual agent believes work promotes productivity it reasons that most agents believe work promotes productivity. Third, behaviour is interconnected: actions do not stand alone, but are similar and influence each other [235]. We find that interconnectivity is reflected in models that use plans, but current agent models do not model explicit relations between activities, between each activity and each other model concept (e,g., desires, needs, resources, locations) nor model hierarchies of activities. In summary, although current agent models capture some aspects of SPT, none fully captures any of the individual aspects, nor is there a model that aims to integrate all three empirical and theoretically grounded aspects.

This chapter shows the usefulness of a computational agent model that inte-

grates SPT and provides requirements that help modellers to achieve this model. We do not argue that the resulting model will provide a general theory of human behaviour, but that integrating SPT is useful for agent-based modellers because it enables new insights using mechanisms that are based on evidence. Although it would certainly be useful to give an exact scope of SPT and its relevance when compared to other theories, this is rather difficult: SPT is a high-level abstract theory, social practices theorists define a SP in different ways, SPT is ever-expanding and merges with other theories and the jury is still out on the so-claimed limitations of SPT [226]. As shown by [226], SPT has given insights in a wide and diverse range of domains: eating, Nordic walking, teaching, learning, washing machine use, cycling, mobility, day trading on the Nasdaq market, domestic energy use, household waste, sustainable design, sustainable consumption, temporalities of consumption, the work of ambulance paramedics or lawyers, anxiety, memory, communities of practice, and organizational learning and knowing. An agent model that integrates SPT will enable researchers to use ABS to gain insights regarding habituality, sociality and interconnectivity in a wide range of social systems.

The remainder of this chapter is structured as follows. Section 3.2 distils requirements for modelling habitual, social and interconnected behaviour from the literature on SPT, agent theory and social psychology. Section 3.3 provides an overview of current agent models and review to what extent they satisfy our requirements. Section 3.4 discusses the consequences of the limitations of current agent models, Turing-completeness and the need for integration of these aspects versus a reductionist scientific strategy. Section 3.5 concludes the chapter.[1]

3.2. Distilling Requirements from Literature

Habituality, sociality and interconnectivity express that SPs have similar properties in three dimensions: over time, over people and over different activities (see Figure 3.1). Habitually expresses that SP are similar over time. For example, commuting is habitual, because a person commutes by car everyday. Sociality express that behaviour is similar over people. For example, commuting is social, because most people associate commuting with a car. Interconnectivity expresses that behaviour is similar for different activities. For example, commuting is interconnected, because most people associate both shopping and commuting with a car. SPs thus truly capture our everyday doings and saying: behaviour that is expected to follow a predictable pattern as its similar with respect to time, people or other activities.

To connect SPT to ABM we need to connect the collective view of SPT and the individual view of agents. Shove et al. [236] sees SPT as a way to abstract away from the individual. They see SPs as a collective entity that recruits or loses host

[1] A note to reviewers. A pre-print of (part) of this chapter is available as [Mercuur, Rijk, Virginia Dignum, and Catholijn M. Jonker. "Modelling Agents Endowed with Social Practices: Static Aspects." arXiv preprint arXiv:1811.10981 (2018)]. This version differs significantly in that it focuses on current literature: it includes a review of current agent models and excludes a proposed model. In addition, the chapter has been thoroughly rewritten to increase clarity, motivation and relevance. If accepted, we will follow the IEEE guidelines by updating the Arxiv version to refer to this chapter and clarify the differences.

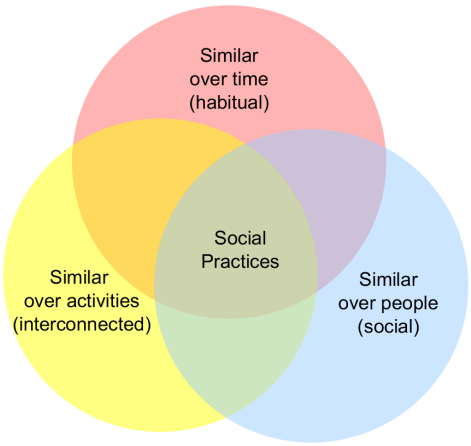


Figure 3.1: A Venn diagram representing how habitual, social and interconnected actions come together in social practices.

(i.e. agents) over time. In contrast, Giddens [103] views SP as a way to connect agency and social structures. In the same line, Dignum and Dignum [77] brings agency back on the table and connects collective concepts related to SPs (e.g., values, norms) with individual agent concepts (e.g., goals, beliefs). For ABM it is in particular important to connect the individual and collective view, because ABM uses agents as a primary concept and aims to connect the micro (individual) with the macro (collective).

To extract requirements for integrating SPT and agent models, the following subsections discuss in more detail how SPs relate to habits, sociality and interconnectedness. Each subsection connects the collective view of SPs to the individual view of agents. We end each subsection with a number of requirements for integrating SPs in ABM.

3.2.1. Social Practices and Habits

SPs and habits both describe behaviour that is similar over time. SPT studies repetitive behaviour on a collective level and is interested in what aspects of a SP exactly repeat [34, 235]. Social psychologists study repetitive behaviour from the individual perspective as 'habits'. They use the term 'habit' to refer to a phenomenon whereby behaviour persists because it has become an automatic response to a particular, regularly encountered, context [154].

We differentiate between two views on habits: habits as a behavioural dynamic notion and habits as a cognitive static notion. The switch from the behavioural view to the cognitive view entails that habits are not merely observable behaviour, but also mental phenomena. For example, one can refer to the habitual behaviour car-driving or the habitual cognitive connection between commuting and car-driving. The static view entails that habits are not only the repeated behaviour over time but also a mental configuration for that behaviour that persists in the mind for

some time, irrespective of the times when the behaviour is actually carried out. For example, one can express a habit dynamically as 'to use the car everyday' or statically as 'at this moment there is a strong mental connection between the car and commuting'. An agent model that integrates SPT thus needs not only a representation of the dynamic decision, update and reasoning algorithms, but it also needs to provide the static configuration of objects, variables and relations. Thus, the model should provide researchers with the primary concepts and relations to model agents that make habitual decisions, updates and reason about habituality.

Models that aim to express habitual decisions and updates need to contrast these with intentional decisions and updates[95, 250, 267]. We follow Wood and Moors by recognizing that the automaticity of habits entails unintentional, uncontrollable, goal independent, autonomous, purely stimulus-driven, unconscious, efficient, and fast behaviour [186, 267]. The automaticity of habits gains meaning when contrasted with another decision mode: intentional decisions. Furthermore, habits and intentions interact and habitual decisions and updates are a product of this interaction [267]. For example, a car-driving habit emerges when agents intentionally drive cars over a long enough period of time. To model the automaticity of habits and interaction with non-habitual behaviour, the model should enable agents to differentiate between habits and intentions.

Habits are sensitive to contextual triggers. For example, the context 'home' can trigger the habit of taking the car (whereas the context 'hotel' might not). To be more precise: habitual decisions are triggered – not by context but – by context-elements [267]. For example, it is not so much home that triggers taking the car, but a combination of context-elements such as being at home, in the morning, together with your partner. The strength of these 'context element'-action relations is a continuous instead of a discrete parameter (e.g., a coffee machine is a slightly stronger habitual trigger than a colleague) [186]. Thus, to form the basis of habitual decisions, the model should enable habitual relations between an action and a context-element where the strength of that relationship is a continuous parameter.

The literature on habits differs in what they consider as context-elements. The common factor in these definitions is that context-elements are physical tangible resources or locations [155, 255, 267]. For example, nearby cigarettes can trigger a habit of smoking. Wider definitions of context-elements allow timepoints, other activities [255] and/or other people to trigger habits Wood and Neal [267]. We choose the wider definition because it matches how habituality is described in SPT. Reckwitz [214] emphasized that habituality in SP does not only refer to physical resources but to mental associations as well. Thus the model should capture that context-elements can comprise resources, activities, locations, timepoints or other people.

Habits depend on their actors. First, the strength of the habitual connection between context and actions is agent-specific. Bob's habit to take the train is not the same as Alice's habit to take the train. Second, how the strength changes over time differs per human. Lally et al. [155] empirically studied this strength gain in an experiment where subjects were asked to do the same action daily in the same context and report on automaticity. The subjects reported an increase in habit

strength that followed a different asymptotic curve per subject and converged at a different maximum habit strength per subject. The model thus needs to enable agents to differ in the strength of the habitual connection, the maximum of this strength and the function over time to reach this maximum.

A habit is sensitive to the attention attributed to a decision [197]. The more attention attributed to the decision the lower the chance the action is done out of habit. The literature on the regulation of attention is extensive (e.g., see [10, 220] for an overview) and it goes too far to capture this concept in detail in this chapter. Enough to say here is that the model should capture that agents can vary in their attention at different moments in time.

Intentional actions contrast with habitual behaviour as intentional actions are attempts to achieve some abstract aim [150]. Examples of concepts that capture this abstract aim are goals, desires, values, motives [125, 150]. The model should provide a concept that captures the abstract aim intentions are directed at.

The following requirements summarize this section:

H.1 The model should capture the similarity of behaviour over time.

H.2 The model should provide researchers with the primary concepts and relations to model agents that make habitual decisions, updates and reason about habituality.

H.3 The model should provide researchers with the primary concepts and relations to model agents that differentiate between habits and intentions.

H.4 The model should support habitual relations between an action and a context-element where the strength of that relationship is a continuous parameter.

H.5 The model should capture that context-elements can comprise resources, activities, location, timepoints or other people.

H.6 The model should capture that agents can differ in

(a) the strength of a habitual connection

(b) the maximum strength of a habitual connection

(c) the time to reach this maximum

(d) the amount of attention they attribute to a decision

H.7 The model should provide a concept that captures the abstract aim intentions are directed at.

3.2.2. Social Practices and Sociality

SPs and sociality are connected because both focus on the similarity of behaviour over people. SPT focuses on sociality primarily as a static group notion emphasizing that we have a similar view on the world that can be organized in terms of our SPs. Social intelligence focuses on sociality as a dynamic individual notion emphasizing

our ability to act wisely in interactions. This chapter uses the literature on SPs and social intelligence to identify what is required to model that SPs are social.

There is a variety of definitions on what it means for an agent to be social or socially intelligent. For Thorndike its the ability to act wisely in interactions [249]. This is close to the layman idea of sociality: an activity that is done in the presence of other people. For Goleman, it means that agents have social awareness and social influence [111]. For Dignum and Dignum [77] its the ability to form expectations about the behaviour of others and react to them. The commonality in these views is that there is some information to be had about other people and that this information is used to guide (social) decisions and (social) updates. The model should provide primary concepts and relations that capture this social information and enable agents to make socially intelligent decisions, update social information and reason about collective concepts.

In the agent literature, there has been an evolvement about which primary concepts should be used to do this social decision-making, updating and reasoning. A first series of chapters uses agents that only take into consideration the actions of others. For example, in the Consumat model [133] an agent takes into consideration what most other agents do. Castelfranchi [43] emphasized that we need to extend such models to also consider the mental state of other agents. He claims the notion of social action cannot be a behavioural notion - just based on an external description, because what makes the action social is that it is based on certain mental states. A second series of chapters focuses on such a representation of the mind of other agents based on individual notions such as beliefs, desires and intentions [38, 88]. Sociality is introduced as a secondary notion. For example, as the ability to form beliefs about other's goals [88] or as a mechanism to filter its intention to a socially desired set [38]. Dignum et al. [71], Hofstede [125] argued that humans are at the core social beings and thus use social concepts as a primary concept. Castelfranchi [45] agrees and sees these social concepts as the "new micro-foundations" for agent decision-making. We view social concepts here as referring to reasoning concepts that depend on being collective (and henceforth call them 'collective concepts'). For example, the notion of culture is hard to imagine without multiple agents. A third series of chapters focuses on these collective concepts. For example, values[51, 174], norms [68], trust [140] and culture [126]. What characterizes this work is to use a collective notion in the individual reasoning of an agent. For Gilbert [106], it's exactly this ability to reason about collective concepts (e.g., culture, or a political party) as individuals that makes us unique as humans. Furthermore, the use of collective notion in individual reasoning provides ABM with a way to connect the micro and macro. We require that an agent model of SPs uses a comprehensive set of collective concepts that support social decision-making, updating and reasoning.

In SPT, an SP is seen as social exactly because it is a concept that depends on being collective: they are the primary concepts that we use to order the world around us. We order our day by a series of practices (breakfast, commuting, working, lunching, sports, showering, sleeping) and we have a similar view on these practices (e.g., to commute with the car one needs a car, believe car driving en-

ables commuting and value going from A to B). Reckwitz [214] emphasized that SPT entails a paradigm shift: the social is captured in our collective SPs instead of in a collective mental world (i.e., mentalism) or collective texts (i.e., textualism). For Shove et al. [235] and Reckwitz [214] an SP is thus social in the sense that it stores social, collective and similar information and not necessarily in the sense that it's interactive. For example, one of Shove et al. [235]'s canonical examples of an SP is showering, an SP that is mostly done alone. Furthermore, SPs capture a particular view on the social world where the social world is ordered around our daily doings and sayings. We require that an agent model of SPs captures that SPs relate the collective social world with the individual world of interactions.

Whereas SPT makes a point of using SP to take a collective view on behaviour that does not concern individual interactions, we see SPs as a primary concept that individuals use to guide interactions. For an SP-theorist such as Reckwitz [214], SPs are thus something collective, "but not in the sense of a mere sum of the content of single minds, but in a time-space transcending non-subjective way". Although SPs thus emerge from individual enactments, Reckwitz [214] views it as a completely separate collective entity. Shove et al. [235] argued that what makes SPT valuable is, in particular, this shift from the individual view to the collective view. Individuals are merely carries or host that are used by the SP to spread around. Shove et al. [235] sees SPT as a way to break with the view that behaviour is the result of individual decision-making. In contrast, for an agent theorist such as Dignum and Dignum [77], SPs are an entity that (also) exist on the level of individual decision-making. As mentioned before, Dignum and Dignum [77], Narasimhan et al. [191] see SP as useful for individual interactions just because it exists on both levels: the individual and the collective level. For agent-based modelling, in particular, it's important to follow this second line of reasoning and include both the individual (micro) and collective (macro) view in our understanding of the social world. SPs are one way to model social information around a structure that exists both the individual and collective level: practices. We require that an agent model of SPs supports agents that order social information around their practices.

By looking at SP from both an individual and collective viewpoint, we notice that these two views do not always match. In other words, there are multiple views possible on the same SP. For example, one can view car-driving as an action that promotes (the value of) pleasure or that demotes pleasure. Moreover, humans differ between what they believe an SP comprises and what they believe others believe an SP comprises. From the viewpoint of the individual, one can differ between a personal view and a collective view on an SP. For example, one believes that car-driving is usually seen as a means for transport but believes him or herself it's a fun activity. These personal and collective aspects of a view can differ: one can believe in a personal view on something without believing this view is collective or one can believe something is the collective view without believing in this view. We require that agents have beliefs about both their individual view on an SP as well as a collective view on an SP.

There is a difference between the collective view on an SP from the viewpoint of the modeller and the viewpoint of the agent. The modeller can see what beliefs are

truly collective among all agents: the modeller can extract the collective SP from a model. For example, a modeller might extract that there are individually different beliefs about the relation between car-driving and pleasure, but that all individuals believe that to car-drive one needs a car. In contrast, the agents themselves have to guess what others believe and these guesses differ. For example, one agent believes that most others see car-driving as pleasurable while other agents believe there are individual differences. There is a reciprocal dynamic between the beliefs that are truly collective among all agents and the collective view of each individual agent.[2] A modeller is able to extract the true collective view from the fine-grained collective views of agents, but not vice-versa. Therefore we require that an agent model of SPs supports a collective view that can differ from agent to agent.

This view also makes clear that SPT only considers a subset of social intelligence. SPT takes a general collective view on our daily activities; SPs are a heuristic where humans generalize over a group of people. This contrasts with another strand of social intelligence called the theory of mind (ToM). Studies on ToM study human's ability to create a mental model of others' beliefs. [269]. In contrast with SPT, studies on ToM consider beliefs about specific others and chains of beliefs. For example, one can belief John believes car-driving is fun or belief that John beliefs that I believe that John believes car driving is fun. These aspects are out of the scope of SPT. SPs are a heuristic that considers only two agents: itself or the group. For example, when greeting someone in most cases it suffices to know that most people view greeting as polite and see shaking hands as a part of greeting. SP focuses on the social intelligence that works in most situations, in contrast, research on the theory of mind treats particular cases where more in-depth reasoning is needed. We thus require that an agent model of SPs supports both a personal and collective view on SP (but not necessarily beliefs about particular others or chains of beliefs).

The following requirements summarize this section:

SI.1 The model should capture the similarity of behaviour over people.

SI.2 The model should provide researchers with the primary concepts and relations to model agents that make socially intelligent decisions, updates and reason about collective concepts.

SI.3 The model should use a comprehensive set of collective concepts that support social decision-making, updating and reasoning.

SI.4 The model should enable agents to order social information around their practices.

SI.5 The model should capture that social practices relate the collective social world with the individual world of interactions.

SI.6 The model should capture that agents have a personal view on a social practice.

[2] And another reciprocal relation between the individual view of the agent and the collective view.

SI.7 The model should capture that agents have a collective view on a social practice.

SI.8 The model should enable agents to have a different personal view than their collective view.

SI.9 The model should enable agents to each have a different collective view on a social practice.

3.2.3. Social Practices and Interconnectivity

SPs and interconnectivity are connected because they both focus on the similarity of behaviour over different activities. Activities here refer to bodily movements. SPT focusses on interconnectivity on an abstract level of activity. For example, it discusses how the work-commute and school-commute are connected because they both use the car as a resource. The agent literature focuses on interconnectivity on a concrete level of activity. For example, it discusses how commuting comprises getting the car keys, driving the car and arriving at work. SPs and agent activities thus exist at different levels of abstraction, but both comprise bodily movements. As we will discuss in the next section, we view agent activities as part of SPs. This section uses the literature on agents and SPs to identify what is required to model that SPs are interconnected.

SPT argues that if SPs are connected in some aspects, then they become more connected in other aspects too [235]. For example, Shove et al. [235] mentions how in the early days of driving, cars easily broke down. To be able to drive a car one needed the competence to repair it. The SP of driving thus became connected with other SPs that related to the competence of repairing, for instance, plumbing or carpeting. The meaning of these SPs as something masculine influenced the meaning of car driving. Car driving and plumbing now share a masculine meaning. The model should provide researchers with the primary concepts and relations to model that if SPs are connected, they become more connected.

In the agent literature, activities are interconnected to enable agents to make decisions and inferences. Hindriks [123] connects activities via goals in the agent GOAL language. For example, the goal of a successful day can be split up in you being in the car, you being at work and you being home. Hindriks [123] views goals as states of the world and activities as ways to reach the state. In contrast, research on language protocols [153] uses activities as their primary concepts and specifies relations between them. A common factor in this work is that it's necessary to define the relation between activities to enable agents to decide what to do next and reason about the properties of activities. Recent work in SPT reflects this view. For example, Cass and Faulconbridge [42] found that interviewees decide to take the car, because they aim to connect leisure activities, healthcare activities and shopping activities. The model should provide the primary concepts and relations to enable agents to make interconnected decisions, updates and reason about the interconnectedness of activities,

In SPT, SPs are interconnected in terms of time, space or common elements [235]. First, SPs connect when they are enacted at the same time or in sequence.

For example, the SP of breakfast and commuting are interconnected because they happen around the same time. Second, SPs connect when they are enacted in the same space. For example, the SP of working and getting coffee are connected because they are happening in the same place. Third, SPs are connected when they share an element. For example, in the early mentioned example of Shove et al. [235] plumbing and car-driving are connected via the competence of repairing and the meaning of masculinity. The model should express that SPs are connected in terms of time, space and common elements.

From the agent perspective, Chen et al. [46], Okeyo et al. [201], Storf et al. [245], van Kasteren et al. [253] classified activities and studied in which possible ways activities relate. The central aim of this work is to recognize activities in the context of smart homes. For example, to recognize that a person is cooking, because he or she boils water and is looking for a cutting board. Okeyo et al. [201] makes a difference between actions and sequential activities.[3] Actions are atomic. Sequential activities are an ordered sequence of actions. For example, commuting is a sequence of taking the kids to school and going to work. Note that from this point on we will separate between activities and actions. Activities refer to any bodily movement (i.e., actions and sequential activities). Actions refer to the subset of activities that are atomic. The model should differentiate between different types of activities: atomic actions and sequential activities (an ordered sequence of actions).

Okeyo et al. [201] separate two types of relations between activities: an ontological and temporal relation.[4] The ontological part describes relations between actions such as subsumptions, equivalence or disjointness. For example, taking the train to school is a kind of commuting. A temporal relation encodes qualitative information regarding time. For example, the user performs two activities after another. This ontological and temporal information can be used by agents to decide what to do next or to make inferences. For example, humans infer that if taking the car to work is environmentally unfriendly then taking the car to school might be as well. The model should capture these temporal and ontological relations between activities to enable agents to make decisions and inferences.

The following requirements summarize this section:

I.1 The model should capture the similarity of behaviour over different activities.

I.2 The model should provide researchers with the primary concepts and relations to model agents that make interconnected decisions, updates and reason about the interconnectedness of activities.

[3] Okeyo et al. [201] identify two other types: simple activities and multi-task activities. These types of activities enable a precise temporal activity model where, for example, activities overlap. However, modelling these type of activities requires a complex quantitative temporal specification that is not needed for the longer temporal scale at which ABS studies systems.

[4] Note that other authors use the term ontology to refer to any kind of relation between two objects. Temporal relations are thus a subset of ontological relations. However, Okeyo et al. [201] uses the term ontological to refer to inferences one can easily make in description logic, whereas he uses the term temporal to refer to relations he can make in Allan's temporal logic.

I.3 The model should express that social practices are connected in terms of time, space and common elements.

I.4 The model should differentiate between different types of activities: atomic actions and sequential activities (an ordered sequence of actions).

I.5 The model should capture both temporal and ontological relations between activities to enable agents to make decisions and inferences.

3.3. Evaluation of Current Agent Models

We evaluate for 11 domain-independent agent-based models to what extent they satisfy the requirements for integrating SPT in agent models. These models are not designed to satisfy our requirements, they have their own purposes. However, the comparison makes clear that if one wants to integrate SPs in agent-based models current models do not suffice. To select related models, we use the overview by Balke and Gilbert[19], but omit neuro-cognitive models because they study the neurology of the mind as viewed from the outside, while we are interested in a socio-cognitive view on the mind as we experience it from the inside. We add two relevant domain-independent agent-based models published after the review of [19]: MAIA[101] and Agent-0[83].[5] MAIA is relevant as it aims to integrate sociaility with ABM and adds a new concept of roles. Agent-0 is relevant as it aims to integrate three decision-making modules that include a context-dependent module and a social module. Using this method we come to a list of the following 11 domain-independent agent models: PRS[96], BDI[213], eBDI[206], BOID[38], BRIDGE[75], EMIL-A[13], NOA[152], MAIA[101], Consumat[133, 135], PECS[252] and Agent-0[83].

We divide these agent-based models into three categories: reasoning models, normative models and social-psychological models. Reasoning models first emphasized autonomy, reactivity (e.g., PRS) and later added proactivity (e.g., BDI, eBDI). When it became clear that adding agent-communication languages to such models does not suffice to successfully represent sociality in humans, researchers focused on adding norms to agent models[41, 70]. Normative agent models focus on different types of norms: social norms (e.g., EMIL-A), deontological norms (e.g., BOID) or both (e.g., MAIA); and different dynamics within norms: norm innovation (e.g., EMIL-A) or norm enforcement (e.g., BOID). Last, some researchers took their inspiration more directly from social-psychological literature and combined several socio-psychological mechanisms in one model (i.e., Consumat, PECS, Agent-0). Reasoning models, normative models, social-psychological models thus represent three categories in ABM that relate in a similar way to our requirements.

The remainder of this section compares the reasoning, normative, and social psychological models to our requirements on habituality, sociality and interconnectivity. If there are differences between the models within the category, then we

[5] We do not focus on agent-programming languages or agent communication protocols, but on conceptual or formal models of agent decision-making.

follow the charitable principle; the comment in the table refers to the model(s) that relate(s) most closely to our requirement and the(se) model(s) are stated.

3.3.1. Habits

Table 3.1 shows that current agent models do not support (1) explicit reasoning about habits, (2) context-dependent habits and (3) individual learning concerning habits. In more detail:

H1-2

Current agent models do not support explicit reasoning about habits. Reasoning agents and normative agents do not make explicit habitual decision and updates or reason about habituality. Norms differ from habits in that they refer to similarity over people (e.g., most people usually drive a car), whereas habits refer to similarity over time for one individual (e.g., I usually drive a car). Consumat models habitual decisions by giving the agent a chance to repeat past behaviour Jager [133]. However, there is no explicit variable capturing the properties of habits (e.g., the strength of the habit). Therefore, the Consumat agent makes habitual decisions but is not able to reason about habits. In summary, some agent models make habitual decisions (Consumat), but none reason about habits.

H3

Only the Consumat agent has both a habitual and intentional decision mode. However, habitual decision-making in the Consumat is not context-dependent (H4-H5) or agent-specific (H6). Reasoning models and normative models aim for reactivity instead of habituality. Reactivity matches habituality in that it requires agent models to react to the environment (i.e., context) in a timely fashion [270]. Reactivity (in current implementations) differs from habituality in that it confounds pre-conditions and triggers (H4), does not consider agents and activities to be context-elements (H5) and is not adaptive or agent-specific (H6).[6] All agent models have an intentional mode of decision-making (sometimes called deliberate decision-making or rational decision-making) (H7). In summary, current agent models all have an intentional mode of decision-making, but only the Consumat also has a habitual decision mode (which has several limitations we will now expand on).

H4

Some agent models conceptualize context-action relations, but they confound pre-conditions and triggers. Reasoning models and normative models confound pre-conditions and habitual triggers[77]. Pre-conditions (or as [77] calls them: affordances) relate context-elements to when an action is possible, whereas habitual triggers relate context-elements to priorities over actions. Confounding pre-conditions and triggers leads to unintentional prioritization in reasoning and nor-

[6] As Balke et al. [19] mentioned, reactivity is modelled on the assumption behaviour is optimal. Habituality differs in that habits are a heuristic: they are a fast automatic response that works in most cases but can be contra-intentional.

mative models [77].[7] In Agent-0 one of the three decision-making modules (the affective one) uses context to determine the appropriate fear reaction. However, this differs from habits where the conditioning happens directly between the context and action. Although the Consumat agent has a habitual decision-making mode, the habit is not sensitive to the context. For example, it is not relevant if the agent is at home or in the office to repeat past behaviour (instead repetition depends on the current satisfaction of the Consumat agent). In summary, reasoning models and normative confound pre-conditions and triggers and only one social-psychological model (Agent-0) supports context-action relations, but only in one module.

H5
Current agent models do not consider other activities, timepoints or agents to be context-elements. Although we found no formal definitions as specifications are example-based, instances of context-element in the models refer only to resources or locations.

H6
Current agent models do not model adaptive and agent-specific reactions to the context. Reasoning and normative models hard-code the reaction to the environment (at all times and for all agents) in the form of plans. Agents thus do not learn an agent-specific reaction to the environment over time. In the Consumat model, the experience of an agent influences its propensity to repeat past behaviour. However, there are no agent-specific parameters that specify an agent's personal tendency to go into a habit or develop stronger habits over time. In summary, in some agent models (Consumat) habits depend on experience, but in none of the agent models, the agents have a personal tendency to go into habits or develop stronger habits over time.

H7
Current agent models all have intentions that are direct at an abstract aim. In reasoning and most normative models intention is a primary concept. In normative models, intentions are captured by utility (MAIA), goals (EMIL-A) or normative goals (EMIL-A). In social-psychological models, intentions are captured by utility-maximization. Current agent models thus capture intentions either as a primary concept or reframe it as goals or utility-maximization.

3.3.2. Sociality

Table 3.2 shows that current models do not use a comprehensive set of collective concepts, order information around actions and relate individual and collective concepts in order to guide interactions. In more detail:

[7] In addition to confounding pre-conditions and habitual triggers, reasoning and normative models do not have a many-to-many relation between actions and context-elements, which is necessary to express how strongly a context-element triggers an action.

S1
Only normative models have an explicit concept that denotes a similarity over people. Norms denote the similarity of actions over people (e.g., most people drive a car). However, normative models do not conceptualize the similarity of people concerning other mental constructs in the model. For example, no concept expresses that most people have a certain goal or that most people relate a certain action to a certain goal. In reasoning models, agents enact the same actions, goals or plans, but there is no explicit concept capturing this similarity. Models for the theory of mind and mental models enhance the architecture of BDI agents (see e.g., Felli et al. [88], Jonker et al. [141]). Although these approaches model some aspects of the collective social world they are not systematically build-up from collective concepts such as values, culture, norms or social practices. This comment is in line with [125]. Likewise, in social psychological models, agents enact similar actions or have similar motivational concepts (e.g., needs in Consumat or fears in Agent-0), but they do not have explicit concepts capturing similarities and dissimilarities and these agents do not reason about such similarities. In summary, reasoning and social-psychological models have no explicit concept that denotes the similarity of agents; normative models capture the similarity of people with respect to actions as a characteristic of the concept norm.

S2
There are aspects of social intelligence that fall outside the scope of this chapter, therefore we withhold from a complete evaluation on the ability of models to do social decision, updates and reasoning. Recall that this chapter looks at the intersection of social practices and social intelligence (section 3.2.2) and aspects like ToM or social impact are not required of a social practice model. Having said that, different agent models emphasize sociality to a different extent. This results in differences in how explicit these models incorporate social decisions, updates and reasoning: the reasoning models do not emphasize explicit sociality, the normative models emphasize sociality wrt norms and the social-psychological models emphasize social mechanisms more than explicit collective concepts and reasoning. The details of these differences are covered in the following sub-sub sections that treat S3-S9. In summary, we state the differences concerning S2 that are relevant for this chapter in the next sub-sub sections but withhold from a complete evaluation of the models concerning S2.

S3
Current agent models do not use a comprehensive set of collective concepts. For example, there is no concept of values, culture or identity. Reasoning models only use individual concepts: beliefs, desires and intentions and eBDI adds one collective concept to this list: emotions. Normative models focus on the concept of norms and MAIA uses the concept of role to denote the subset of agents that are similar in their goals or norms. Social-psychological models use emotions (Agent-0) and social mechanisms (imitation), but no explicit collective concepts. In summary, the current agent models use some collective concepts (norms, emotions, roles), but omit others (values, culture and identity).

S4

Current agent models do not order sociality around practices. Reasoning models, normative models and social-psychological models do not conceptualize practices but do have a concept of action. Because practices are a series of actions, we instead inquire: do current agent models order their information around these actions? More precisely, do current agent models conceptualize the relation between each action and each other concept within the model and use these relations to understand each other? We find that they do not. In reasoning and normative models, actions, desires and intentions are linked via plans. This does not specify for every action if the action promotes a desire or not. In social-psychological models, agents conceptualize the relation between actions and mental concepts (e.g., needs) and physical concepts (i.e., in the fear-module of Agent-0), but not other actions. Because actions (and their relation with other concepts within the model) do not take center stage, they are not used to form mental models of other agents (i.e., order social information). In summary, as agents do not use actions and practices as a central component of their model these cannot be used to order social information.

S5

Through comparison with current models, we found that requirement SI.5 needs to be evaluated on two aspects. Requirement SI.5 states that the model should capture that social practices relate the collective social world with the individual world of interaction. This encompasses two aspects. First, sociality requires that the agent model connects collective concepts to individual concepts (SI.5a). Second, sociality requires that agents use collective concepts to guide interactions (SI.5b). We now continue to evaluate current agent models on both these aspects.

S5a

Current models do not relate individual concepts to collective concepts, because they do not comprise both the individual and collective concepts or because the ontology does not relate them adequately. First, all of the evaluated models lack collective concepts. In particular, values, culture or identity. Second, all models lack individual concepts. In particular, habits (except, as discussed, the Consumat model). Habits form the individual counterpart to norms. Where habits state what an individual mostly does, given a certain situation, norms state what the collective mostly does, given a certain situation. None of the models (including normative models) relate habits explicitly to norms. Third, for some models, the individual and collective concept is included but not adequately connected. We give two examples (in these examples the collective and individual concept have the same name):

- all models do have a concept of the physical world, which is both an individual and collective concept, but do not reason about the fact that the physical world is collective. The models do not reason about the fact that others, given the current physical context, will do the same action as they do, or have the same

desire (BDI), same emotion (agent-0) or same need (Consumat).[8].

- none of the models reasons about the fact that the motivational constructs they use are collective. The models do not reason about the fact that most other agents have the same desires (reasoning and normative models), emotions or needs as themselves.

In summary, current models do not relate individual concepts to collective concepts, because they do not comprise certain individual concepts (i.e., contextualized habits) and collective concepts ((i.e., values, culture or identity) or because the ontology does not relate them adequately (i.e., physical context, desires, emotions, needs).

S5b

Current agent models are limited in guiding interactions because they do not relate individual and collective concepts. The different cases of limited individual and collective connection (S5a) have a direct consequence for guiding interactions. First, if an agent model does not comprise a collective concept (e.g., culture), the agent cannot form expectations of other agents that are based on that concept (e.g., the other agent does not shake hands because that's not part of its culture). Second, if a model does not comprise an individual concept (e.g., habits), the agent cannot reason about the collectiveness of this individual concept (e.g., most other people will also have the habit to drive to work, so I can ask someone to carpool with). Third, if a model does not make an adequate connection between an individual and collective concept (e.g., the individual and collective view on motivation), the agent cannot use its mental model to form expectations about others (e.g., most other people will also have a desire to be healthy).[9] In summary, in current agent models agents are limited in guiding interactions, because they cannot form expectations regarding collective concepts they do not conceptualize or use their mental model as a proxy for others.

S6-9

Normative and social-psychological models agents have a different personal and collective view on actions. In these models, agent each have a different view on the best action based on their intentions (i.e., intentions or utility-maximization). Besides, they have a different view on what the collective views as the best action. In BOID and BRIDGE, this collective view is expressed in the different obligations that hold for different agents. In MAIA [101], this collective view is expressed in roles, where a different role for an agent means a different conceptualization of the norm. In Consumat and Agent-0, this collective view expressed in the form

[8]The only exception being the MAIA model where agents reason about the collectively of physicality through contextualized norms: other agents mostly do X given physical context Y

[9]This limitation connects to [71, 72, 125] who state that sociality should be ingrained in the core of their reasoning. Instead of adding extra concepts to model sociality (e.g., by extending BDI models), the same concepts should be used to form individual reasoning as well in forming expectations about others. As such, sociality is not added as a layer on top of individual reasoning but is used to shape the reasoning of the agent.

of a local norm: agents each imitate the agents around them and thus have a different view on what the collective best action holds. Because none of the models connects other individual and collective concepts, the agents do not have a different personal and collective view on these concepts.[10] In summary, in normative and social-psychological models each agent has a different personal and collective view on actions, but not on other concepts.

3.3.3. Interconnectivity

Table 3.3 shows that current models do not make explicit relations between activities, between each activities and each other model concept (e,g., desires, context) nor model hierarchies of activities. Reasoning and some normative models (BOID, BRIDGE) do have plans that implicitly connect activities and context-elements and motivations. Social-psychological models connect activities to the same motivations. However, without explicit hierarchies and connections between activities the models do not directly support inferences about the similarity of activities on a personal level (e.g., two activities need the same resource, need to be performed in sequence, promote the same value) or on a social level (e.g., other people will need this resource). In more detail:

I1

Current models do not model an explicit similarity relation between each activity. Reasoning and normative models use plans to relate activities to other activities. The relations between activities and the similarities in activities are implicit in these plans. For example, two activities lead to the same goal, because the two activities lead to two different subgoals that eventually lead to the same goal (see [225] for BDI-based planning). Social psychological models model relate actions with motivational constructs in the model. For example, a certain action satisfies a certain need (Consumat), fear or rational intention (Agent-0). In summary, reasoning and normative models specify a relation between activities via plans, all models specify relations between activities and some other elements, but there are no explicit relations between activities that denote the similarity.

I2

Current models focus on deciding what's next, some models (extensions of BDI) plan ahead, but none reason about the interconnectivity of activities. Reasoning and normative models use plans to make decisions and reason about the interconnection of activities. BDI-based models in particular reason about what is the next action. Hunsberger [130], Planken et al. [208] extend such models by planning ahead: they approach the interconnection of actions as a coordination problem where agents search optimal sequences of activities that satisfy a set of time constraints. This is useful for ABS that zoom in on a small time-scale (where precise coordination

[10] Reasoning models that aim at representing others' mind focus on social expectations about *specific* others (e.g., [88, 212]), whereas SPT emphasizes a heuristic humans use where they focus on *the* others. Humans make assumptions about how most others view an activity [69]. We require that models differ between themselves and 'the group'.

of actions with other agents has a big influence on the overall system). However, ABS aims to model human limitations in sequencing actions: action sequences in humans are the sub-optimal product habits and, coordination between humans is based on the limited beliefs agents have about others. Reasoning and normative models do not emphasize making inferences using the interconnection of activities to further resource management or social expectations. For example, agents do not reason that 'car commuting to school' relates to 'car commuting' and therefore cannot infer that the resource 'car' relates to the abstract activity of 'commuting'. As a consequence, the agent does not immedatly know that it will need a car to commute nor is the agent able to form expectations about others needing a car to commute (also see Section 3.4 about the interplay of sociality and interconnectivity). Social-psychological models focus on deciding what is next, but do not plan ahead nor reason about the interconnectivity of models. In summary, all models focus on deciding what's next, some extensions of BDI models focus on optimal plans, but none emphasize reasoning about the interconnectivity of activities.

I3

Current reasoning and normative model link temporal, spatial and elemental through plans. Social-psychological models connect activities directly to motivational constructs. Agent-0 makes a spatial connection between activities in the fear-module by making the fear an agent experiences context-dependent. All other explicit temporal and/or spatial relationships have to be specified by the designer of the system. However, there are no further spatial and no temporal connections in the model. In summary, in reasoning and normative models the connection of activities (temporal, spatial, to other elements) is implicit in plans; in social-psychological models, actions are connected directly to motivational constructs, but there are no temporal and limited spatial (only Agent-0 and only in one module) connections.

I4

Current agent models do not have different types of activities. BDI-based models (eBDI, BOID, BRIDGE) are centered around states of the world: goals/desires are states of the world and agents have beliefs about these states being currently true or not true. For example, an agent reasons that it wants to go from state `on(block-a, floor)` to `on(block-a, block-b)` and therefore first moves block-b on the table and then block-a on block-b, but does not have explicit knowledge about the sequence of activities it should take. Social-psychological models emphasize the mental models of agents modelled around motivational construct (e.g., an agent has a disposition towards this need or this fear), but do not focus on hierarchies of activities. In summary, current models do not have different types of activities, because they take states of the world or agents – and not activities - as the primary concepts of their reasoning.

I5

Current models have implicit (reasoning and normative), limited or no (social-psychological) temporal and ontological relations between activities. As mentioned by now, reasoning and normative models make connections between activities via

plans, but these are not explicit. For example, there are no statements about 'car commuting' being a kind-of 'commuting' or 'bringing your kids to school' being a part-of 'commuting'. Social-psychological models make no temporal or ontological relation between activities.

3.4. Discussion

The agent models we reviewed are all Turing-complete: they are expressive enough to simulate the computational aspects of any other real-world general-purpose computer or computer language. However, we did not evaluate the models on their ability to make certain calculations, but if the concepts SPT emphasizes are primary concepts and relations in the model. This entails that the model expresses the semantics of the concept: it models the meaning of the concept by specifying the relation with other objects and imposing certain restrictions on the allowed deductions. Thus we make a difference between between enabling in the wider sense of Turing-complete and enabling as a primary concept. A good example to illustrate this is the use of habits. Reasoning models enable (in the wider sense) the modeller to simulate habitual behaviour in an agent: a series of very-specific plans ensures that the agent keeps repeating the behaviour in a certain context. However, they do not enable (in the narrow sense we use) the modeller to specify and interpret habitual behaviour. Likewise, the Consumat model enables (in the wider sense) the modeller to simulate content-dependent habits: a series of very specific needs and actions ensure the agents only repeat behaviour in a certain context. However, it does not specify the semantics of habits that detail habits are context-dependent. In both cases, the models would become difficult to manage and interpret if one wants to analyze habits. In summary, we are not interested enabling modellers to make certain computations, but in enabling modellers to express the semantics of habits, sociality and interconnectedness and integrating these aspects via primary concepts and relations in the model.

This chapter shows that current agent models do not support (1) explicit reasoning about habits, (2) context-dependent habits and (3) individual learning concerning habits. As shown in Section 3.2.1, both in SPT and social psychology habits are recognized as a key component of behaviour. By not supporting explicit reasoning about habits, agents are not able to correctly combine habits with other decision-making concepts. For example, an agent is not able to reproduce the ability of humans to put itself intentionally in a context (e.g., the desk) to trigger a habit (e.g., to work) [267]. Without modelling how habits depend on context the activities of agents will repeat an activity in any context. For example, an agent is not able to reproduce the behaviour of a person that habitually drinks coffee at work, but tea at home. Without modelling individual learning concerning habits an agent is not able to acquire new personal habits or lose old ones. For example, agents are not able to reproduce differences in humans where one person gets easily stuck in the habit of car-driving, while another person switches between driving a car and using a train. Concluding, new agent models need to be developed to integrate social-psychological research on habits in agent models.

This chapter shows that current models do not use a comprehensive set of

collective concepts, order information around actions and relate individual and collective concepts in order to guide interactions. As shown in Section 3.2.2, both in SPT and agent theory sociality is recognized as a key component of behaviour. Without integrating a comprehensive set of collective concepts agents cannot fully reproduce human ability to reason about a collective world. For example, without concepts such as values, culture and identity an agent cannot understand why the other agent refuses to shake hands and propose a solution. Although practices (i.e., actions) are not the only concept around which social information can be ordered, ordering information around practices has at least two advantages: a practice is social, that is, it exists on both the individual and collective level (see Section 3.2.2) and ordering social information around practices corresponds to empirical work in neurology on social reasoning Metzinger and Gallese [178]. Without connecting individual and collective concepts agent cannot extend their reasoning about their individual preferences to forming expectations or preferences of others. For example, an agent cannot reason that because the agent itself has a habit to drive a car, chances are high that others share this habit and therefore ask a colleague to carpool. Concluding, new agent models need to be developed to integrate research on SPT and social agents to model sociality in agent models.

This chapter shows that current models do not make explicit relations between activities, between each activity and each other model concept (e,g., desires, context) nor model hierarchies of activities. As shown in Section 3.2.3, in SPT, interconnectivity is a key component of behaviour and, in agent theory, interconnectivity is gaining evidence as a useful way of modelling decisions. Without explicit hierarchies and connections between activities, current models do not directly support inferences about the similarity of activities on a personal level (e.g., two activities need the same resource, need to be performed in sequence, promote the same value) or on a social level (e.g., other people will need this resource). Concluding, new agent models need to be developed to integrate SPT and agent research on interconnectivity in agent models.

To integrate SPT in ABM, we need to integrate habituality, sociality and interconnectivity in one agent model. This chapter follows a reductionistic approach by splitting up SPT in aspects and these aspects in requirements. This approach has been highly successful in the physical sciences and makes it possible to understand complex systems by understanding the properties of presumably more basic components. However, it gives the false impression that investigating the organizational features of things is less informative than investigating component properties [58].[11] For example, it is the integration of habits and interconnectedness that enables a model to express the routine of first taking the kids to school and then going to work. It is the integration of sociality and interconnectivity that enables agents to infer that others expect that you take the car as a way to commute. Thus although some of the agent models perform relatively well when evaluated against a single aspect or a single requirement, to truly integrate SPT in agent models we

[11] (Neuroscience is a good example of where a nearly complete theory of synaptic function and only a slightly less complete understanding of neurons has led to a less dramatic understanding of human behaviour or social systems than one envisioned [58].)

need to integrate habituality, sociality in interconnectivity in one model.

3.5. Conclusion

This chapter provided a set of requirements for integrating SPT in agent models. We identified three empirically and theoretically relevant aspects of SPT for modelling agent decision-making: habituality, sociality and interconnectivity (Figure 3.1). Section 3.2 discussed these aspects using literature on SPT, agent theory and social psychology and provided a list of requirements for an agent model that aims to integrate SPT.

This chapter provided an evaluation of 11 current agent models against the requirements we elicited. We found that current agent models do not fully capture habituality, sociality or interconnectivity nor is there a model that aims to integrate all three aspects. First, current agent models do not support (1) explicit reasoning about habits, (2) context-dependent habits and (3) individual learning concerning habits (see Table 3.1). Second, current models do not use a comprehensive set of collective concepts, order information around actions and relate individual and collective concepts in order to guide interactions (see Table 3.2). Third, current models do not make explicit relations between activities, between each activity and each other model concept (e,g., desires, context) nor model hierarchies of activities (see Table 3.3). In addition to detailing these specific differences, we discussed that to utilize SPT in ABM, we need to integrate habituality, sociality and interconnectivity in one agent model. Concluding, although all agent models capture some aspects of SPT, none fully captures any of the individual aspects, nor is there a model that aims to integrate all three empirical and theoretically grounded aspects.

This chapter shows the usefulness of a computational agent model that integrates SPT and provides requirements that help modellers to achieve this model. As we discussed, all the agent models we review are all Turing-complete, but they do not incorporate these aspects as first-principles. Therefore, modellers are not supported in modelling habits, sociality and interconnectivity. We discussed several examples of human behaviour that current domain-independent agent models do not support. First, without an adequate model of habits, an agent is not able to reproduce the ability of humans to put itself intentionally in a context (e.g., the desk) to trigger a habit (e.g., to work). Second, without an adequate model of sociality, an agent cannot reason that because an agent itself has a habit to drive a car, chances are high that others share this habit and therefore ask a colleague to carpool. Third, without an adequate model of interconnectivity, an agent cannot reason that it will need to use the car for commuting to school, because commuting to school is a kind of commuting. Last, without integrating habituality, sociality and interconnectivity in one model, agent models do not support agents that combine habits and interconnectivity in a routine of first taking the kids to school and then going to work. Furthermore, the model does not support agents that, in addition, use sociality to expect others have a similar commuting routine and therefore decide to carpool together. Integrating SPT in agent models will help us understand the world in terms of three key aspects of behaviour: lazily habitual, lovingly social and actively interlinked.

Table 3.1: =
Habituality: Verification and Explanation of Different Agent Models with Respect to our Requirements.
The '?'-column denotes weather the requirement is unsatisfied (✗), somewhat satisfied (∼) or satisfied (✓) by the agent model.

Requirements		**Reasoning Models** (PRS, BDI, eBDI)		**Normative Models** (BOID, EMIL-A, NOA, MAIA)		**Socio-Psych Models** (Consumat, PECS, Agent-0)	
Nr.	Shorthand	?	Explanation	?	Explanation	?	Explanation
H1-2	similarity over time captured in habits	✗	no explicit habitual reasoning	✗	norms do not refer to individual repetition	∼	habitual decisions, but no reasoning about habits
H3	habits and intentions	∼	no, but reactivity and deliberation	∼	no, but reactivity and deliberation	✓	habits and intentions (Consumat)
H4	action-'context-element' relation	∼	no difference between pre-condition and trigger	∼	no difference between pre-condition and trigger	∼	only in one module (Agent-0)
H5	a context-element comprises	✗	activities, time and agents are not context-elements	✗	activities, time and agents are not context-elements	✗	activities, time and agents are not context-elements
H6	agents differ in their habits	✗	reactivity is not adaptive nor agent-specific	✗	reactivity is not adaptive nor agent-specific	∼	only due to experience
H7	intentions are directed at an abstract aim	✓	intentions are a primary concept	✓	captured by intentions, utility, goals or normative goals	✓	intentions are captured by utility-maximization

Table 3.2: Sociality: Verification and Explanation of Different Agent Models with Respect to our Requirements. The '?'-column denotes weather the requirement is unsatisfied (✗), somewhat satisfied (~) or satisfied (✓) by the agent model.

Requirements		**Reasoning Models** (PRS, BDI, eBDI)		**Normative Models** (BOID, EMIL-A, NOA, MAIA)		**Socio-Psych Models** (Consumat, PECS, Agent-0	
Nr.	Shorthand	?	Explanation	?	Explanation	?	Explanation
S1	conceptualizes similarity over people	✗	only implicit in similar actions, goals or plans	✓	norms explicitly capture similarity wrt actions	✗	only implicit in similar actions or motivations
S2	social decisions, updates and reasoning	~	no explicit sociality	~	explicit decision, updates and reasoning about norms	~	imitation of other agents, but no social reasoning
S3	use collective concepts	✗	only emotions (eBDI)	~	only roles (MAIA) and norms	✗	only emotions (Agent-0) and implicit norms
S4	order sociality around practices	✗	mental models are not ordered around actions	✗	mental models are not ordered around actions	✗	mental models are not ordered around actions
S5a	relate collective and individual concepts	✗	limited individual/collective concepts and relation	~	only relates personal actions and collective actions	✗	limited individual/collective concepts and relation
S5b	use collective concepts to guide interactions	✗	no, because of limitations in S5a	~	does not adequately relate norms to habits in interactions	✗	no, because of limitations in S5a
S6-9	different personal and collective view	✗	no collective view	✓	agents each track a local norm wrt actions	✓	agents each track a local norm wrt actions

Table 3.3: Interconnectivity: Verification and Explanation of Different Agent Models with Respect to our Requirements. The '?'-column denotes weather the requirement is unsatisfied (✗), somewhat satisfied (~) or satisfied (✓) by the agent model.

Requirements		**Reasoning Models** (PRS, BDI, eBDI)		**Normative Models** (BOID, EMIL-A, NOA, MAIA)		**Socio-Psych Models** (Consumat, PECS, Agent-0	
Nr.	Shorthand	?	Explanation	?	Explanation	?	Explanation
I1	similar over activities	✗	only implicit similarity in plans	✗	only implicit similarity in plans (BOID, BRIDGE) and norms	✗	only implicit similarity in needs (Consumat) or emotions and rationality (Agent-0)
I2	interconnected decisions, updates and reasoning	~	use plans to decide what next and plan ahead	~	use plans (BOID, BRIDGE) and norms to decide what is next, but no reasoning	✗	no planning or reasoning
I3	SPs are connected in time, space and common elements	~	only implicit connection through plans	~	only implicit connection through plans and norms	~	connected to needs/emotions, but limited spatial and no temporal connections
I4	different types of activities	✗	no separation between different levels of abstractness in activities	✗	no separation between different levels of abstractness in activities	✗	no separation between different levels of abstractness in activities
I5	temporal and ontological connections	~	implicit connection through plans	~	implicit connection through plans	✗	no temporal and ontological relations

4

Conceptualising Social Practice Theory for Agent-Based Models

This chapter is in submission at the Journal of Artificial Societies and Social Simulation.

The model of the Homo economicus explains action by having recourse to individual purposes, intentions and interests; social order is then a product of the combination of single interests. The model of the homo sociologicus explains action by pointing to collective norms and values, i.e. to rules which express a social 'ought'; social order is then guaranteed by a normative consensus. In contrast, the newness of the cultural theories [RAM: e.g., social practice theory] consists in explaining and understanding actions by reconstructing the symbolic structures of knowledge which enable and constrain the agents to interpret the world according to certain forms and to behave in corresponding ways. Social order then does not appear as a product of compliance of mutual normative expectations, but embedded in collective cognitive and symbolic structures, in a 'shared knowledge' which enables a socially shared way of ascribing meaning to the world.

A. Reckwitz, Toward a theory of social practices: A development in culturalist theorizing

4.1. Introduction

Our routines play an important role in a wide range of social challenges such as climate change, disease outbreaks and coordinating staff and patients in a hospital. Studying these systems via agent-based simulations enables researchers to gain insight into complex aspects of these challenges such as human interaction, individual adaptability, heterogeneity, feedback loops and emergence. To use agent-based simulations to understand the role of our routines in social challenges we need an agent framework that integrates our routines [176]. Mercuur et al. [176] recognizes this and argues for grounding such an agent framework in social practice theory (SPT). Social practice theory is a socio-cognitive theory that emphasizes how humans use their routines (practices) to come to a common view of the world (a social view). Using a socio-cognitive theory, such as SPT, enables ABS researcher to reuse evidence from the social sciences and relate the agent framework to other social theories. In short, translating SPT to an agent framework supports researchers in simulation studies on our routines that are grounded in evidence and embedded in the social sciences.

To further an agent framework that integrates SPT, Mercuur et al. [176] identifies relevant aspects of SPT and establishes a set of requirements for integrating SPT in agent-based models (ABM). They identify that SPT gives insight into the habitual, social and interconnected nature of our routines (see Figure 4.1). For example, the SP of commuting consists of a series of interconnected actions that are chained habitually: getting in the car, driving the kids to school and getting to work. These practices – chains of actions – are used in social situation as a mental model to reason about others actions (e.g., to coordinate the carpool). The authors distill requirements for integrating these aspects from the literature on agent theory, social psychology and SPT. For example, a requirement related to this example states that agents should use social practices as a collective view on the world to

coordinate their actions (see Appendix 8.3 for all requirements). The authors evaluate current agent frameworks and conclude that – although these frameworks are useful for their own purposes – they do not satisfy the requirements for modelling our routines. In short, a new framework is needed that integrates SPT in agent models and satisfies the requirements set out by Mercuur et al. [176].

This paper extends current work that integrates SPT in agent models. Narasimhan et al. [191] used SPT in an ABS to study a specific domain: energy systems. In contrast, our paper emphasizes reusability by providing a *systematic* translation from theory to framework to enable a *domain-independent* framework. Holtz [127] studies SPT via ABS by studying SPs as if they were agents themselves. In contrast, our work integrates SPT and agent theory enabling insights in the interaction of humans and SPs. Most notably for our purpose is the high-level conceptual model of SPs by Dignum and Dignum [77] and the related formalization in the preprint Dignum [69]. Dignum [69] presents a formalization that integrates SPT and agent theory with a different scope, audience and resulting view on SPs than ours. Dignum [69] nevertheless serves as an essential inspiration and starting point for our framework. As such, we consider the concepts Dignum and Dignum [77] uses in our own framework and relate our framework to theirs in the discussion.

This paper provides the domain-independent social practice agent (SoPrA) framework that satisfies the requirements set out by Mercuur et al. [176]. We approach this by using concepts from the literature on agent theory, social psychology and SPT described in Mercuur et al. [176]. For each modelling choice, we present an overview of the relevant concepts using delineated boxes. We describe how based on these concepts we provide a new conceptualisation (i.e., new labels and a new organization) that fulfils our requirements while aiming for reusability (i.e., compactness and modularity). The concepts from the literature are marked as either (1) directly represented as a concept in the UML (✓), (2) (partly) subsumed by the concepts in our UML (≃) or (3) left out the UML because they are not necessary to fulfill our requirements (✗). Our UML provides constrains in how the chosen classes relate by depicting associations, the multiplicity of the associations and the types of the attributes. For example, our UML shows locations are not also resources as objects cannot be a member of two distinct classes. In short, by building upon concepts of the literature we provide a clear and correct framework in UML.

SoPrA is implemented in UML, OWL and Java. The main body of this paper describes our framework in UML. UML is a standardised, popular language that makes the framework easier read, understand, code, extend, maintain and adapt [25]. We show the UML adheres to the complete set of requirements and correctly implements them. The implementation of the requirements is by design, i.e. each subsection argues for a specific framework that embodies the requirement. Our online repository provides a formal framework in the OWL-language, which shows our framework is consistent, and a computational framework in Repast Java, which enables simulating social phenomena with the framework (see Appendix 8.3). These frameworks provide ABS researchers with reusable, consistent and computational framework to simulate our routines and gain new insights into social systems.

The remainder of this paper is structured as follows. Section 4.2 presents the

high-level choices made to connect SPT and agent theory. Section 4.3-4.5 provides the primary concepts and relations to researchers to model respectively habituality, sociality and interconnectivity. Section 4.6 discusses the resulting framework by exemplifying the new insights enabled via simulation and the criteria on which we made model choices. Furthermore, Section 4.6 concludes that our paper provides ABM researchers with a computer model to simulate our routines and gain new insights into social systems. Our paper ends with three appendices: Appendix 8.3 describes the requirements set out by Mercuur et al. [176], Appendix 8.3 provides possible extensions of SoPrA. Appendix 8.3 describes the computational and formal versions of SoPrA.

Figure 4.1: Social practices describe behaviour that is similar over time, activities and people (adapted from [176]).

4.2. High-Level Modelling Choices

SoPrA connects the collective perspective on behaviour of SPT and the individual perspective of agent theory. To capture the collective perspective of SPT, SoPrA should capture the similarity of behaviour over time, people and activities. To capture the individual perspective of agent theory, SoPrA should provide researchers with the primary concepts and relations to model agents that make (1) habitual decisions, updates and reason about habituality, (2) socially intelligent decisions, updates and reason about collective concepts and (3) interconnected decisions, updates and reason about the interconnectedness of activities (H2, S2, I2). Figure 4.2 provides a high-level overview of the concepts and relations we use to connect these two perspective: activities, elements, agents and activity-associations. The remainder of this section explains how these concepts connect SPT and agent theory.

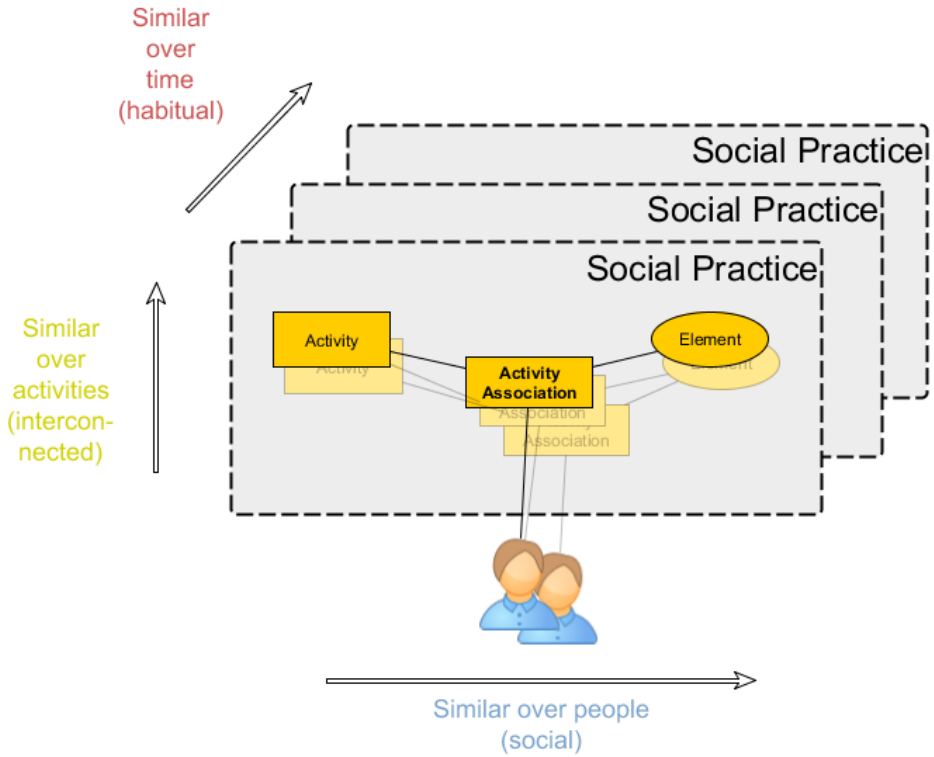


Figure 4.2: A high-level overview of SoPrA depicting the main concept – activities, elements and agents – and the ternary connection between these concepts that can be similar over time, people and activities.

Activities refer to the bodily movement itself without all the mental connotations. In SPT there are two views on the relation between activities and SPs. First,

according to Reckwitz [214] an SP represents a pattern which can be filled out by a multitude of single and often unique activities reproducing the SP. An SP thus consists of several activities. For example, the SP of commuting can consist of taking your kids to school and then going to work. Second, for Shove et al. [235] activities are not a part, but instead another view on the SP. They view the SP as either a collection of elements (named practice-as-entity) or as something that is performed (named practice-as-performance). For example, they view commuting as either a collection of elements, such as the meaning of transport and the material car, or as a series of activities, such as getting in your car and going to work. We follow Reckwitz [214] in that activities are elements that are part of an SP (and not different views on an SP) because this matches the agent perspective where activities are also part of an agent model (and not different views on an agent model). In agent theory, activities (or actions, see 4.5) are the result of a decision-making process. For example, an agent chooses to move a block, commute by car or sell a stock. In short, activities are part of both SPs and agent models and therefore enable us to connect both perspectives.

Elements refer to epistemological entities such as places, human values or resources. In agent theory, these elements are part of the mental model of the agent in the form of typical agent constructs such as goals, desires or resources. In SPT, the aforementioned practice-as-entity perspective entails that SPs are collections of interconnected elements. Taylor [247] first describes the SP as an entity separate from an agent with multiple elements: "meanings and norms implicit in [...] practices are not just in the minds of the actors but are out there in the practices themselves". SPs thus comprise other elements than activities (e.g., meanings and norms). [214] describes the connection between these elements and the SP as that the existence of SPs depends on the existence and specific organization of these elements and cannot be reduced to any single one of them. For example, the SP of skateboarding depends on a specific group of elements that interconnect: materials like the skateboard or street spaces, the competences to ride the board, rules and norms like what defines a trick, and the meaning different groups attribute to the practice like recreation or transport [235]. In short, elements are part of both SPT and agent theory. This paper describes elements described in SPT and agent theory and chooses which elements suffice to conceptualize an agent model that integrates SPT (and reflects the requirements).

Agents refer to the individual decision-makers that represent humans. Whereas in agent theory the agent is a central component in SPT the agent is subsidiary. In SPT, the social practice 'recruits' agents as 'hosts' and uses them as a vehicle to spread [236]. In fact, one common motivation to use SPT is to be able to abstract away from the actor [214, 227, 236]. This paper acknowledges the strength of SPT in expressing this collective perspective, but to serve ABM aims to connect this perspective with the individual perspective of agent theory.

Activity associations refer to the mental connection between an activity an element. In agent theory, agents make mental connections between activities and other mental constructs. For example, BDI-agents connect their desire D to activity A by having the belief that A satisfies D. In SPT, activity associations are called con-

nections; they connect an activity and another element of an SP [34, 214, 235]. In SoPrA, activity associations are the central component that connects SPT and agent theory. Figure 4.2 shows how activity associations connect activities, elements and agents.

SoPrA associates elements and agents (e.g., meanings, materials) with specific activities and not — as is common in SPT literature [34, 214, 235] — with the SP as a whole. For example, an agent does not associate the SP of commuting with environmentalism. Instead, an agent associates an activity instantiating commuting — taking the train to work — with environmentalism, while associating another instance— taking the car to work — with the meaning of efficiency. We give four reasons for this modelling choice. First, this enables modelers to capture the beliefs an agent has about an SP in a nuanced way (e.g., make a difference between forms of commuting that promote different values). Second, it enables modellers to use these nuances to guide the decision-making of the agent (e.g., choose car-driving over the train because it's more efficient). Third, it enables modellers to feed SoPrA with a selection of facts about actions and let the agent make inferences about composite or sequential activities (e.g., commuting is boring, because all of its implementations are boring). Fourth, it corresponds to how empirical work in neurology conceptualizes human decision-making [178]. In Metzinger and Gallese [178], associations between activities and goals (i.e., one of the SP elements) are used both to reason about our own decisions as well as a way to reason about others (i.e., social intelligence). In short, associating elements with specific activities – instead of with the SP as a whole – enables modellers to capture the nuances within a SP, use these nuances to guide decision-making, use a circumscribed knowledge base and corresponds to evidence within neurology.

Figure 4.2 presents how our requirements relate to the choices made in conceptualising an agent model that integrates SPT: a focus on activities as part of a SP, conceptualising SPs as a collection of elements and emphasizing the associations between activities and elements. These choices enable us to express more clearly what it means for behaviour to be similar over time, people and activities (requirement 1, 4, 6). The similarity of behaviour lies in activity associations. Thus habituality captures that people associate the same SP elements with an activity over time; sociality captures that different people associate the same elements with an activity; interconnectivity captures that people associate the same elements with different activities. SPs thus capture that activity associations are similar over time, people and different activities.

This leaves open the question of how we make these concepts and relations precise to enable agents to reason about habits, sociality and interconnectivity (H2, S2, I2). Figure 4.3 presents an overview of the full UML model with the specific concepts and relations that suffice to implement the requirements. The remainder of this paper refers to this figure and explains why these concepts and relations model habitual, social and interconnected behaviour.

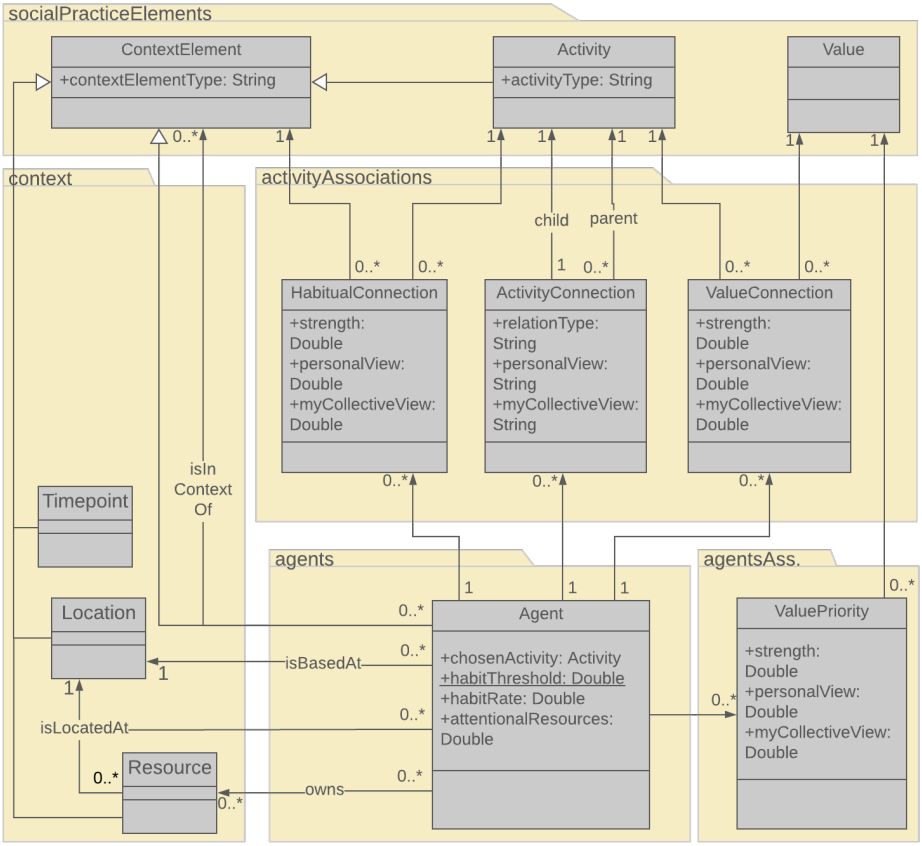


Figure 4.3: SoPrA in UML that integrates social practice theory and agent theory emphasizing habituality, sociality and interconnectivity.

4.3. Conceptualising Habituality

SoPrA provides the primary concepts and relations to researchers to model habitual decisions, updates and reasoning (H.2). The more concrete requirements on habits (H3-H7) specify in detail what is needed to model habits. For example, agents need to be able to differentiate between habits and intentions (H3). Figure 4.3 shows that agents are able to differentiate between a habitual association (i.e., `HabitualConnection`) and an intentional association (i.e., `ValueConnection`). This leaves open how we conceptualize what triggers habits (H4), what concept captures habitual triggers (H5), how we enable agents to differ in their habits (H6) and why intentions are directed at an abstract value (H7). The following subsections treat these aspects in more detail.

4.3.1. Context-Elements

Our conceptualisation of context-elements captures that context-elements can comprise resources, activities, location, timepoints and other people (H.5). Aside, the conceptualisation of context-elements should use comprehensive set of collective concepts needed for social reasoning (S3).[1] Box 4.3.1 gives an overview of the concepts relevant to conceptualize the `context` package that fulfills requirement H.5 and S.3. The remainder of this subsection explains how we used these concepts to construct SoPrA.

Relevant Concepts 4.3.1

Material$^{\simeq}$ Materials refer to the physical aspects of a SP [157, 214, 227, 235] . For Shove et al. [235] this comprises the resources people relate to the SP. For example, the SP of commuting relates to a car, bike, or public transportation.

Context$^{\simeq}$] Context is any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between the user and the application, including the user and the applications themselves [64].

Context Cue$^{\checkmark}$ A contextual cue is any event noticed by the organism, with the exclusion of the target stimuli that form the learning experience (Learning and Memory: A Comprehensive Reference, 2008)

Location$^{\checkmark}$ A location is a fixed geographical point.

Resource$^{\checkmark}$ A resource describes a physical entity that has some utility.

Time-Point$^{\checkmark}$ A fixed temporal point. For example, morning, afternoon, 12pm or Wednesday.

To satisfy requirement H.5 we model a `ContextElement` class that generalizes the `Activity` class, `Agent` class, `Timepoint` class, the `Location` class and the `Resource` class. The UML clarifies the relation between these classes (e.g., a resource cannot also be a location as UML forces these objects in seperate classes). The `ContextElement` class has an attribute `contextElementType` that can be assigned to the string `"Activity"`, `"Agent"`, `"Location"`, `"Resource"` or `"Timepoint"` such that an agent can easily differentiate between the types of context-elements. We chose the wording context-element instead of context cue, because these objects have more function in the model than 'cueing' a habit. The concept of 'material' is entailed by the concept of 'context-element' (in particular, by the class `Resource` it generalizes) and therefore does not require a separate class. Locations, resources, timepoints, agents and activities each trigger habitual

[1] Physical objects namely play a central role in our social world in that they provide a collective common ground [232].

behaviour, but have additional, but different, functions in the model and therefore are modellled as seperate classes. This conceptualisation of context-elements thus captures all necessary elements that trigger habits and still enables these elements to have different functions in the model.

SoPrA models context-elements as tokens – not types. The type-token distinction refers to the difference between a general sort of thing and its concrete instances [266]. For example, the token `Car1` is a possible context-element, but the type `Car` is not. Usually a Class in UML refers to a type and an instance of the class to a token. However, because we aim for a domain-independent framework we abstract over multiple types of context-elements (cars, restaurants, homes) for which different tokens trigger different habits. For example, a modeller should separate between `BobsCar` and `AlicesCar`, the former triggering a habit for Bob while the latter does not. Note that the car of someone else, to some extent, also triggers your car driving habit. Modellers capture this by modelling a context-hierarchy that relates a more abstract context-element to a more particular one. For example, to model a class 'Car' that has multiple instances 'Bobscar' and 'AlicesCar'. SoPrA thus models context-elements as tokens to provide the fine-grained static groundwork on which, depending on the target system, researchers can build more complex context hierarchies.

4.3.2. Habitual Connections

SoPrA supports habitual connections between an action and a context-element where the strength of that relationship is a continuous parameter (H.4). Box 4.3.2 gives an overview of the concepts relevant to conceptualize the `HabitualConnection` class that fulfills requirement H.4. The remainder of this subsection explains how we used these concepts to construct SoPrA.

Relevant Concepts 4.3.2

Connection$^{\checkmark}$ The relation between the elements of a social practice. For example, the connection between a commuting activity, material car and the value of environmentalism.

Habitual Trigger$^{\simeq}$ The cue in the context that triggers the habit [192].

Habit Strength$^{\simeq}$ The extent to which an action is experienced as habitual. There are several methods for measuring habit strength: the response frequency measure (RFM) [257], behavioural frequency [268] and the self-reported habit index (SRHI) [256].

The `HabitualConnection` class associates an `Activity`, `ContextElement` and an `Agent`. The class captures the mental connection between a context-element, an activity and an agent. In contrast to a habitual trigger it denotes the relation and not the context-element itself. The `strength` attribute of `HabitualConnection` class represents to what extent the context element is connected

with the activity. The stronger these associations the more chance these context-elements will trigger a habit. In social psychology, habit strength is measured on different levels of specificity: per activity (across different context) (RFM[257] and the SRHI [256]) or per activity and per context [268]. In SPT, a connection relates a general activity to a specific element. SoPrA provides the fine-grained groundwork underlying each of these conceptualisations: we denote the habitual strength of a mental connection between a specific activity (e.g., commuting by car instead of the more general commuting) and a specific context-element (e.g., my front door instead of the more general my home).

4.3.3. Heterogeneity in Agents' Habitual Tendencies

SoPrA supports modelling agents that differ in (1) the strength of a habitual connection, (2) the maximum strength of a habitual connection, (3) the time to reach this maximum and (4) the amount of attention they attribute to a decision (H.6). Box 4.3.3 gives an overview of the concepts relevant to conceptualize the `Agent` class that fulfills requirement H.6. The remainder of this subsection explains how we used these concepts to construct SoPrA.

Relevant Concepts 4.3.3

Attention$^{\simeq}$ The allocation of cognitive resources among ongoing processes [11].

Attentional Resources$^{\checkmark}$ The limited amount of cognitive resources available at one moment of time [11]. [11].

The `Agent` class contains an `attentionalResources` attribute that captures the current amount of attentional resources available to allocate to a decision. The amount of attention attributed to a decision is the product of a number of cognitive processes using the same attentional resources [11]. In line with this thought, we chose to model the `attentionalResources` attribute as an integer.

The `Agent` class contains a `habitRate` attribute that ensures heterogeneity in how fast agents acquire habits. This attribute captures how much the `strength` of the `HabitualConnection` class increases when an agent experiences an action in relation to a `ContextElement`. As shown in [170], a difference in `habitRate` suffices to acquire heterogeneity in both the learning rate as well as the maximum habit strength agents report.

The `Agent` class contains a `habitThreshold` attribute that captures a cut-off point: if the agent experiences an habitual pressure above the habit threshold it will make a habitual decision instead of an intentional decision. The habit threshold is the same for every agent as `habitRate` and `attentionalResources` suffice to acquire heterogeneity in the inclination of agents to fall into a habit.

4.3.4. A Teleological Construct

SoPrA provides a concept that captures the abstract aim agents direct their intentions at (H.7). Box 4.3.4 gives an overview of the concepts relevant to conceptualize the `Value`, `ValuePriority`, `ValueConnection` class that fulfills requirement H.7. The remainder of this subsection explains how we used these concepts to construct SoPrA.

Relevant Concepts 4.3.4

Meaning$^{\simeq}$ Symbolic meanings, ideas and aspirations associated with a SP such as, attributing the meaning of 'health' to the SP of riding the bike [235].

Goal$^{\times}$ A goal is an idea of the future or desired result that an agent commits to achieve [159].

Desire$^{\times}$ Represent the motivational state of the system [213].

Value$^{\checkmark}$ Represent what one finds important in life (Poel & Royakkers 2011). For example, privacy, wealth or fairness.

Social Motives$^{\times}$ Basic natural incentives that give rise to "energizing" subsequent action . In paricular, (1) achievement, (2) power, (3) affiliation and (4) avoidance [166].

Need$^{\times}$ Represents what motivates and drives humans. For Maslow [165] needs are more than what is necessary for human survival. The chase to satisfy needs makes "man is a perpetually wanting animal." Maslow [165] represents five needs arranged in a hierarchy of prepotency. These are physiological, safety, love, esteem, and self-actualization.

The `Value` class represents the teleological concept agents' intentions are aimed at. We chose (human) values to model meaning because this concept is (1) teleological, (2) operationalized, (3) shared and (4) trans-situational.

First, values can also be defined as 'ideals worth pursuing' [60]. Values are ends that people what to achieve: they are unreachable, but can be pursued [265]. Values thus give reasons for agents to choose an action that contrasts with their habitual tendency. For example, by attaching the meaning efficiency to car-driving we can contrast the decision of an agent to take the car out of habit with the decision to intentionally take the train because it promotes environmentalism. Values thus provide the abstract aim intentions are directed at.

Second, Schwartz [229] developed several instruments (e.g. surveys) to operationalize values. Values thus provide an operational concept that has been made measurable and understandable in terms of empirical observations. This contrasts with the concept of 'meaning' in SPT, where different authors wrote about different aspects of meaning. Schatzki [227] emphasized the 'teleoaffective' mental elements, such as embracing ends, purposes, projects and emotions. Reckwitz

[214] focused on mental elements that are emotional or motivational. Shove et al. [235] uses 'meaning' as a bucket term that in addition comprises norms, ideas and aspirations.

Third, the categorization of values by [229] has been extensively empirically tested and shown to be consistent across 82 nations representing various age, cultural and religious groups [26, 56, 92, 229, 230]. Values thus provide a shared concept that has a similar meanings to people around the world. For example, by relating car-driving with efficiency one gives it an aim that is understood in cultures all around the world. This satisfies our requirement to model a comprehensive set of social concepts.

Fourth, values are trans-situational [265]. Values abstract over multiple situations and actions. This allows us to interconnect actions from different domains. For example, environmentalism can relate both to the practice of commuting as well as dining. This serves agent-based modellers that model individuals as part of a larger social systems over longer periods over time and, as such, comprise multiple actions that need to be related. Values contrasts with goals that focus on more concrete states within a domain and as such serve more precise co-ordination between agents. For example, car-commuting and hiking promote the same value (e.g., environmentalism), but not to the same goal (e.g., arriving at work, being active).

To enable heterogeneity between agents and the actions agents prefer we conceptualize the `ValuePriority` class and `ValueConnection` class. The `ValuePriority` class with an attribute `strength` represents how important an agent finds a value. For example, one agent finds the value of environmentalism more important than another agent. The `ValueConnection` class with the attribute `strength` represents how strongly an agent relates a value to an action. For example, one agent relates the value of fun to skateboarding while the other does not. The agent bases its intentional decision on strength of the `ValueConnection` class and `ValuePriority` class. For example, an agent might value environmentalism highly and relate the activity `ride bike to work` to environmentalism, therefore, it chooses `ride bike to work` over `drive car to work`. The `ValuePriority`, `ValueConnection` and `Value` class support modellers in modelling agents that direct their intentions at an abstract aim.

4.4. Conceptualising Sociality

SoPrA provides researchers with the primary concepts and relations to model agents that can make social decisions, update social variables an that reason about sociality (S2). Central to our conceptualisation of sociality is the connection between individual and social concepts. For example, norms, the concept referring to what the collective usually does, is connected to the concept of habits, the concept referring to what the individual usually does. Individual and social concepts are two sides of the same coin, or more explicit, two views on the same concept: a personal view and a collective view (S6, S7). In terms of views, habits are the personal view on what an agent usually does and norms are the collective view on what an agent usually does. The remainder of this section explains how we conceptualize a personal and collective view, link individual and collective concepts and model an implicit and explicit view. The resulting conceptualisation supports ABMs where agents use social practices to make social decisions such as coordinating their commute with their colleagues or partners.

4.4.1. A Personal and Collective View

SoPrA provides a personal and collective view on a social practice (S.3). Box 4.4.1 provides definitions of the concepts of common belief and collective intention that are relevant to conceptualize the `personalView`, `myCollectiveView` attributes that fulfills requirement S.3.

Relevant Concepts 4.4.1

Common Belief$^{\simeq}$ Everyone believes ϕ if all agents individually belief ϕ. There is a common belief in ϕ if everyone believes in ϕ, and also everyone believes everyone believes in ϕ, etc. [180]

Collective Intention$^{\simeq}$ Collective intentionality characterizes the intentionality that occurs when two or more individuals undertake a task together [36, 231]

SoPrA uses the `personalView` attribute to refer to an agent's personal view on a concept and the `myCollectiveView` to refer to an agent's collective view on a concept. The `personalView` attribute specifies how strongly an agent itself believes an activity and an element are connected. The `myCollectiveView` attribute specifies how strong an agent believes the collective views (i.e, they believe) an activity and an element as connected. For example, in the `HabitualConnection` class the `personalView` attribute expresses how strong an agent believes that itself habitually connects a context element with an activity (e.g., an agent believes itself has a habit to use the car to commute). Analogously, the `myCollectiveView` attribute expresses how strong an agent believes others habitually connect a context element with an activity (e.g., how strong the agent believes that others habitually use their car to commute). The `personalView` and `myCollectiveView` attributes also feature in the `ChildParentRelation` class, `ValueConnection` class and `ValuePriority` class. The type of the view at-

tributes is the same as the type of the main attribute in the class. Thus, in the `HabitualTrigger`, `ValueConnection` and `ValuePriority` class the type of the view attributes is an integer analogue to the `strength` attribute. In the `ChildParentRelation` class the type of the attribute is a string anologue to the `relationType` attribute (see Section 4.5). In short, the personal view and the collective view are captured in the `personalView` attribute and `myCollectiveView` that feature in the `HabitualConnection`, `ChildParentRelation`, `ValueConnection` and `ValuePriority` class.

SoPrA ensures that agents are each able to have a different collective view (S9). This is emphasized by the 'my' in `myCollectiveView`. This differs from the framework introduced in Dignum and Dignum [77] and formalized in Dignum [69]. Dignum [69] conceptualize an SP as *one* collective entity. For example, there is *the* SP of commuting instead of *a* (view on the) SP of commuting. The SP in Dignum [69]'s framework is one that – if agents choose to participate in the SP – everyone believes (and everyone believes that everyone believes (i.e., a common belief, see Box 4.1). As such, Dignum [69]'s conceptualize social practices as entailing a collective intention.[2] Conceptualising the SP as a single collective entity fits problems that focus on short-term action coordination where it's paramount all agents agree on the context (e.g., human-agent interaction). This is indeed one of the applications [69] has in mind [17]). SoPrA enables a fine-grained view where each agent is able to have a different collective view. This enables to model clusters of collective views on SPs (e.g., one social group views biking as a normal form of commuting whereas another does not). This fits problems studied using ABS: ABS researchers simulate the long-term dynamics of different views on the social word and their influence on both individuals and the social system. When necessary, SoPrA is easily adapted to match Dignum [69]'s framework by fixating the collective view (i.e., making the variable `myCollectiveView` static) and as such allowing only one collective view on an SP.In short, SoPrA enables a fine-grained difference between collective views suitable for ABM and is easily extendable to support *one* collective view whenever necessary (e.g., when applied to human-agent interaction).

SoPrA enables agents to have a different personal and collective view on an SP (S8). For example, an agent believes it personally has a habit to commute by skateboard but the collective has a habit to commute by car. This independence between views is conceptualized in the independence of the `personalView` and the `myCollectiveView` attribute. This way of separating the agent's personal view from its collective view differs from Dignum [69]'s framework. In Dignum [69], agents are defined by concepts that are different from the concepts used to define an SP (e.g., a separation between habits and norms, goals and purpose, personal values and collective values). This separation between agents and SPs enables different types of agents to participate in the SP as long as a relation is specified between their concepts and the concepts used in the SP. Dignum thus aims for a meta-framework that enables plugging in agents specified in different languages (e.g., BDI-agents, norm-based agents, Procedural Rule-Based agents). SoPrA aims for parsimony with an accessible computational implementation in ABM

[2] Or in Dignum [69]'s words: a collective context that agents use to coordinate their actions.

and therefore matches the language of the agent directly with the language of the SP. This facilitates both the parsimony of the model and the ability of an agent to make inferences between its personal and collective beliefs (S5). For example, an agent uses its own belief that commuting is usually done by car – its car habit – to infer that other people usually use a car to commute – a car norm – and asks another agent to carpool. SoPrA thus connects individual and social concepts to promote both the parsimony of the model and to supports agents in using the collective social world in individual decision-making.

4.4.2. A Comprehensive Set of Social Concepts

By connecting individual and collective concepts SoPrA includes a comprehensive set of collective concepts (S3). Box 4.4.2 provides the definitions of a comprehensive set of collective concepts (i.e., norms, values, landmarks and purpose) that are either subsumed by our UML or not relevant for our requirements.

Relevant Concepts 4.4.2

Norms$^{\simeq}$ Norms generally refer to what is standard, acceptable or permissible behaviour in a group or society [91].

Values$^{\checkmark}$ Represent what one finds important in life [209]. For example, privacy, wealth or fairness. Schwartz [229] shows values represent labels that are collective (i.e, consistent across 82 nations).

Landmarks$^{\simeq}$ Landmarks are states that represent states in a protocol where agents have a common belief about having reached that state [153].

Purpose$^{\times}$ The purpose of an action is the reason for which the action is performed. The purpose is the common denominator over a sequence of actions and actors. For example, in the context of a lecture several actions (e.g, speaking or raising hands) done by several actors (e.g., students, lecturer) are all done for a common purpose (i.e., the students learn about the topic) [69].

SoPrA captures norms, values and collective activity associations. Norms were already treated as an example throughout this section as the collective counterpart of habits. The `myCollectiveView` attribute in `HabitualConnection` class captures what an agent believes is the collective view on what actions usually occur given a certain context. This provides the basic ingredients for norms (i.e., in the terminology of Ostrom [53] it contains the Actor, Aim and Condition elements of a norm). Values were treated as necessary as the teleological counterpart to habits (see Section 4.3.4). Section 4.3.4 emphasized the personal view on values but values also have a collective meaning. Recall that the universality of values as demonstrated by Schwartz is one reason we chose values over other teleological concepts. SoPrA captures this collective view on values in the `myCollectiveView`

attribute. Activity associations are necessary to conceptualize interconnectivity and agents and also entail a collective view. Section 4.5 treats this in more detail but for now it suffices to understand that a collective view on interconnectivity enables agents to understand how others believe activities are connected (corresponding to Landmarks (see Box 4.4.2). For example, one believes others view car commuting as a kind-of commuting and therefore asks a colleague to carpool. Although rarely made explicit such beliefs are paramount to interpreting activities of others [178, 201]. In short, by including a collective view on the individual concepts necessary to model habituality and interconnectivity SoPrA captures norms, values and collective activity associations.

SoPrA provides a structure that enables a correct inclusion of additional collective concepts. For example, SoPrA provides the structure to correctly define more detailed statements about habits and norms. Currently, the `HabitualConnection` class captures a basic descriptive belief about what actions usually occur given a certain context. The `HabitualConnection` class is extendable by adding a `sanction` attribute. This enables a modeller to specify rules of the kind "An agent may not drive a car in the city centre or else he will be fined" (see the MAIA framework [101]). The structure of SoPrA ensures such a `sanction` attribute entails both a personal view (e.g., I feel guilty when I drive) and a collective view (e.g., the collective fines me when I drive). Another example would be adding the concept of goals (as featured in Dignum [69]'s framework) and its collective counterpart: purpose (see Box). Goals and purpose provide teleological concepts that capture concrete reachable states of the world (as compared to the more abstract values) and are useful for applications of short-term human-agent interaction (as opposed to long-term ABM). Thus, SoPrA ensures that when a collective concept is added an individual counterpart is defined (and vice versa) such as adding 'purpose' requires a 'goal' and a 'sanctioned norm' requires a 'sanctioned habit'.

4.4.3. The Implicit and Explicit View

SoPrA enables modellers to make a difference between the information implicit for the agents versus information explicit to the agent. For example, agents are able to have a strong implicit habit to interrupt people in conversation without having explicit awareness of this habit. As such, SoPrA enables differentiating agents that are able to only habitually decide to interrupt people from agents that are aware of this behaviour and able to change it. SoPrA captures this difference in a `strength` attribute – an implicit strength – and a `personalView` attribute – an explicit habit strength. A difference in agents personal implicit strength and personal explicit view fits ABMs focus on behavioural change (e.g, tipping points only occur when there is an increased awareness of current habits). Besides, the provided structure – that is, a difference between implicit and explicit variables – is extendable to other variables. For example, an implicit collective view and explicit collective view fit applications in human-agent interaction. By making a difference between implicit and explicit information our model provides a structure to differentiate making decisions (based on implicit information) from reasoning (based on explicit information).

This section showed how agents use a different personal and collective view to

come to a set of social concepts that is used for social decision-making, updates and reasoning. This also clarifies the scope for social decision-making, updates and reasoning with social practices. For example, social practices are a heuristic used to make a difference between a personal and collective view and do not focus on chains of beliefs (e.g., I believe John believes I believe…) as represented in the 'theory of mind' theory[269]. The next section treats the third aspect of SPs: interconnectivity.

4.5. Conceptualising Interconnectivity

SoPrA provides researchers with the primary concepts and relations to model agents that can make interconnected decisions, update connections between activities and that reason about the interconnectedness of activities. Activities are connected *indirectly* via time, space and common elements (I.3). These connections are captured in aspects of the model already treated. The `HabitualConnection` class connects multiple activities to the same location or time-points. The `ValueConnection` class connects multiple activities to the same element (i.e., values). In addition, SoPrA enables *direct* connections between activities. These direct connections are featured in this section. They are temporal or ontological (I5) and connect different types of activities (I4). An example of different types of activities and their connections – for the domain of community – is found in Figure 4.4. The activity tree allows agents to, for example, reason that commuting has two parts (bringing the kids to school and going to work), reason that both parts are implementable using a car (i.e., drive car to school and drive car to work), and reason others believe commuting has two parts (and therefore borrow the car to your partner who needs it to bring the kids to school). The remainder of this section will treat this activity hierarchy in more detail using Figure 4.4 as an example. Box 4.5 provides the definitions of abstract actions, landmark states, Allen's logic and plan patterns that are relevant to conceptualize the `ActivityConnection` class that fulfills requirements I.1-I.5.

Relevant Concepts 4.5.0

Landmarks$^{\simeq}$ Landmarks represent states in a protocol where agents have a common belief about having reached that state [153].

Allen's Logic$^{\simeq}$ Allen's logic supports temporal relations that are relative, imprecise and uncertain [7]. Allen's logic does not attribute a specific time t to an event but the relation between two intervals. Give the intervals t and s some of examples are: t before s, t during s, t equal to s or t overlaps with s.

Plan Patterns$^{\simeq}$ Plan patterns describe usual patterns of actions defined by the landmarks that are expected to occur [69].

SoPrA uses the `activityType` attribute in the `Activity` class to make a dif-

ference between atomic activities, sequential activities and abstract activities (I4). Atomic activities represent concrete bodily movements that agents perform. In Figure 4.4, they are at the leaves of the activity tree, for example, `take train to school` is an atomic activity. Sequential activities are sequences of actions. In Figure 4.4, `commuting` is a sequential activity that sequences`bring kids to school` and `go to work`. Abstract activities are placeholders for different implementations of that activity. For example, `bring kids to school` is an abstract activity that is implementable by `take train to school`, `ride bike to school` or `walk to school`. Although abstract actions are not part of our requirement they are a consequence of the subsumption relation described in the next paragraph. These different types of activities enable modellers to model activities on different levels of abstractness and form the first part of the activity hierarchy.

The second part of the hierarchy consists of different connections between activities defined by the `ActivityConnection` class. This class has the attribute `relationType` that specifies two different types of relation between activities: the `isA` and `partOf` relation. The `isA` relation models the ontological subsumption relation between two activities. For example, taking the train to work is a kind of going to work. This is depicted as a white arrowhead in Figure 4.4. The `partOf` relation models a temporal relation between two activities: it specifies that all the child activities need to be completed to complete the parent activity. For example, car-commuting consists of bringing the kids to school and going to work. Additional constructs for relating activities are found in Dignum [69] and Allen's Logic [7]. Some are not required for our model because they suit coordination on a short time-scale (e.g., the parallel performance operator). Others are covered in our model by a combination of constructs. For example, a temporal sequence relation (i.e,. one action precludes the other) emerges (over time) in our model as a habitual connection between two activities instead of being fixed. SoPrA thus uses a `isA` and `partOf` relation to define the temporal and ontological connection necessary for ABM.

These different types of activities and different type of activity connection form an *explicit* activity hierarchies. As mentioned in an earlier paper, this contrasts with BDI planning models where actities are a means to reach a certain state of the world (a goal) and the connections are *implicit* in plans. By emphasizing the explicit relations between activities, we enable agents to reason about how these activities are connected. For example, by explicitly making the connection between bringing kids to school and driving a car to school an agent can reason about how bringing kids to school requires the resource car. Besides, the agent can ask their partner to bring their kids to school as they will have the same view on the connection between car commuting and bringing kids to school. Thus, by modelling explicit activity hierarchies we support both resource management and social reasoning about connections between activities.

Dignum [69]'s framework falls in between BDI agent and SoPrA in that it features activities in two ways: to reach a state as well as to connect activities. Explicit relations between activities feature in Dignum [69]'s framework in plan patterns. These plan patterns sequence abstract action much like in SoPrA, but in contrast

to SoPrA and in line with BDI, Dignum [69] always mentions abstract actions in relation with a goal ϕ that the abstract action aims to achieve. This allows agents to have freedom in how to achieve an abstract action (i.e., how to achieve the goal phi related to the abstract action), while still keeping a certain rigidity by fixing how abstract actions are sequenced (as featured in the plan pattern). In comparison to SoPrA, this enables different types of agents to be plugged in. This is suitable for Dignum [69]'s purpose: the coordination of different agents in human-agent interaction but increases the complexity of their framework. SoPrA is aimed at ABM and therefore supports resource management and social reasoning while simplifying other parts to promote parsimony and clarity.

Having an explicit and complete activity hierarchy provides agents with a structure they can use in making interconnected decisions. For example, the decision-making of an agent starts at the activity at the top of the activity tree and goes step by step down the activity tree until it reaches an atomic activity. At each step, the agent uses the activity associations related to the candidate activities to make a decision. For example, it chooses go to work by car because it promotes efficiency. Or, it habitually chooses `walk to school` because there is a strong habitual trigger between the abstract activity `bring kids to school` and `walk to school` (recall that activities are also context-elements and thus habitually trigger other activities). After completing an action an agent returns either to the top of the activity tree (to make a decision at the next step) or to a sequential activity to complete both parts of the sequence. Or in a model featuring multiple social practices, the agent switches to another activity tree to sequence drinking a cup of coffee while keeping in mind the necessary actions related to commuting. The activity tree thus has a strong influence on the decisions agents make as it prescribes the possible actions and the possible decision paths in taking those actions. Therefore an ABS researcher should model the activity tree not only in accordance with empirical data but also the purpose of the study. For example, our example tree places great emphasis on bringing kids to school and going to work being two parts of commuting. This is useful to study the connection between the school-commute and work-commute but less suitable for other purposes.

4.6. Discussion and Conclusion

This paper provides the domain-independent SoPrA framework that satisfies the requirements set out by [176]. SoPrA provides an integration of habituality, sociality and interconnectivity in one agent model. SoPrA fills the three gaps in current agent models as identified by [176]. First, it provides a framework that supports context-dependent habits, individual learning concerning habits and explicit reasoning about habits. Second, it provides a comprehensive set of collective concepts, orders social information around actions and relates individual and collective concepts in order to support agent interactions. Third, it defines relations between activities, between each activity and each other model concept and enables hierarchies of activities. All of the above allows modellers to simulate the dynamics of our routines. For example, Mercuur [167] used a preliminary version of SoPrA to show that an individual meat-eating habit more easily breaks in a new context (e.g., a

vegetarian barbecue) and that such a habit, in turn, influences the collective view in the old context (i.e., your family having a more moderate view on vegetarians). The simulation shows that this results in a cascade of habit-breaking and, as such, a relatively effective intervention. By integrating habits, sociality and interconnectivity SoPrA extends current agent frameworks and as such enables new insights on our routines in complex systems.

This work builds upon the work by [69] but focuses on an agent framework for ABM instead of an agent framework for human-agent interaction. This is reflected in the choices presented throughout the paper. Most importantly, the choice to enable a connection between each activity, element and agent (Section 4.2), the choice to use habitual triggers as a primary concept in the model (Section 4.3), the choice to enable multiple collective views on an SP (Section 4.4.1) and the choice to use the same concepts to model the agent's personal view and collective view (Section 4.4.2). These choices result in a framework that supports ABM in simulating the long-term dynamics of complex systems. It enables modelling agents with habits and local collective views (e.g., one social group views biking as a normal form of commuting whereas another does not). Besides, by grouping certain concepts and using the same concepts to capture agents and SPs the framework is relatively accessible and parsimonious. For example, the concept of 'actors', 'resources', 'places', 'start condition', 'duration' are all captured in the `ContextElement` class and its relations to other concepts. Accessibility is, in particular, relevant for ABM, where social scientists aim to use computational tools. Parsimony is relevant for agent-based simulations (ABS) as this enables modellers to run more agents (representing larger groups of people) and keep the parameter space manageable. For applications where a more constraining framework is necessary SoPrA is extendable by fixing certain attributes. For example, by making the variable `myCollectiveView` static and as such allowing only one collective view on an SP. Thus, the choices made in this paper result in a parsimonious, accessible and flexible extension of [69] suited for ABM.

SoPrA focusses on modularity to enhance its (re)useability for ABM. Modularization is a central desideratum in software engineering [24, 100, 179, 254]. By composing a system in modules, relatively independent units of functionality, we enhance its understandability and reuse. The modularity of SoPrA is reflected in its separation in classes and relations. This allows modellers to remove or reuse classes to fit the framework to his or her purpose. For example, removing the `ActivityConnection` class and restricting the model to atomic activities results in a model that emphasizes habituality and sociality (but not interconnectivity). Another useful purpose of modularity in ABM is to be able to trace back the results of a simulation to a particular module by switching part of the model on and off. This enables the modeller to check the validity of particular modules and gain clearer explanations via the simulation.

To further enhance the (re)useability, we aimed for compact modules, while at the same time, maintaining a correct, complete and consistent reflection of the requirements. Section 4.3-4.5 have described the complete set of Mercuur et al. [176]'s requirements and shown how they are correctly implemented in SoPrA. Ap-

pendix 8.3 describes our OWL framework that proves the consistency of SoPrA. We achieve the compactness of SoPrA in a process akin abstract painting where one depicts an object in a few on-point strokes of the brush. In modelling, one reorganizes and relabels concepts until the concepts form independent dimensions that correctly represent aspects of human behaviour. SoPrA uses three on-point dimensions: habituality, sociality and interconnectivity. Other complex concepts are reduceable to these dimensions. For example, sequenced behaviour is a combination of habituality and interconnectivity; a series of activities that trigger the next activity via a habitual trigger relation. A norm is a social version of a habitual trigger. A landmark action or plan pattern is a social version of an activation connection. In short, by focusing on three orthogonal dimensions and making grounded choices SoPrA is both correct, complete, consistent and compact.

Future work should focus on static or dynamic extensions of SoPrA. SoPrA is fit for our purpose – modelling routines – and meets our requirements. By connecting SoPrA with other modelling concepts (e.g., affordances, competences, roles, culture) one enables insights in the interaction of routines with other aspects of social systems. We foresee no problem in such an integration due to the modular and orthogonal nature of SoPrA. Appendix 8.3 provides a UML model that integrates SoPrA with additional agent concepts found in the literature and serves as a template for other desired extensions. Another avenue is to model the dynamic aspects of SPT by create a systemic translation from SPT to a domain-independent agent framework. This paper provides the static groundwork for such an extension.

The contribution of this paper is the SoPrA framework that provides ABM researchers with a means to simulate our routines. The language we use in SoPrA illuminates the habitual, social and interconnected aspects of social systems via simulation. This means we enable a new way to know and explore the world and gain new insights. These insights lead to new ways to control and improve our world. The view we enable is one of empirical and theoretical importance. Our routines have been emphasized since the old Greeks and can now be combined with the strengths of computer simulation. We envision that this allows insights into the interplay of habits and norms; for instance, can we understand the slow change in climate behaviour as a consequence of the interplay of group inertia (due to norms) and individual inertia (due to habits)? How to best influence multiple societal groups using the dynamics of individual habits and local norms to our advantage? Are these behaviours intrinsically connected with other behaviours through routines and do they need to be targeted as such; what does changing one of these behaviours mean for the others? SoPrA provides the language to formulate these questions in precise terms. Furthermore, this better formulation is supported by SoPrA that allows simulation dedicated to exploring these questions.

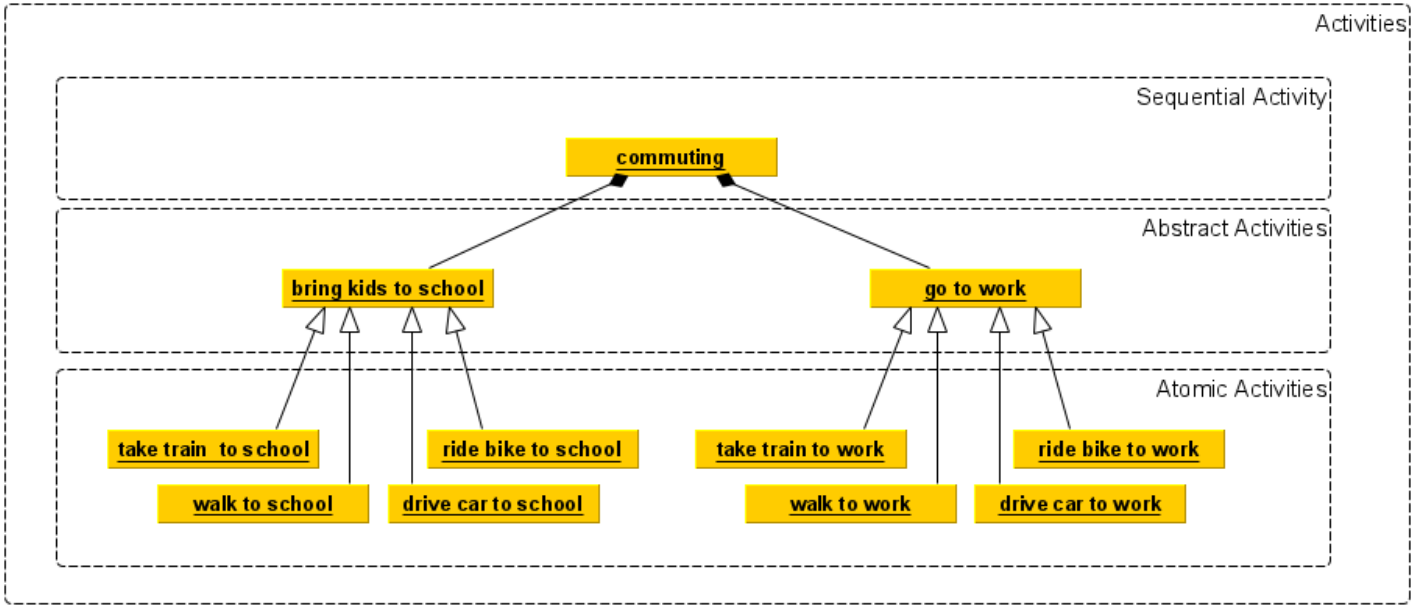


Figure 4.4: An activity tree relating atomic activities, sequential activities and abstract activities for our use case of commuting. The white arrows specify an `ActivityConnection` of the type `isA` and the black diamonds an `ActivityConnection` of the type `partOf`

III

Applications of Agent Framework

5

Aligning Values in AI: Improving Confidence in the Estimation of Values and Norms

5.1. Introduction

As autonomous agents (AAs) become more pervasive in our daily lives, there is a growing need to reduce the risk of undesired impacts on our society [9, 116]. Hence, we need design and engineering approaches that consider the implications of ethically relevant decision-making by machines, understanding the AA as part of a socio-technical system [74, p.48]. For this, we need to ensure that the decisions and actions made by the AA are aligned with the stakeholders' values ("what one finds important in life" [57]) and norms (what is standard, acceptable or permissible behavior in a group or society [90]).

Values and norms can be seen as criteria for decision-making: values relating to more abstract and context-independent ideals, and norms as more concrete context-dependent rules. Furthermore, since humans use values and norms in their explanations, the use of these concepts allows for a better explanation of the AA [161, 174, 184]. People have a common base system of values and they are relatively stable over the life span of a person [52], nevertheless, values and norms vary significantly for each person and socio-cultural environment, making it difficult

This chapter has been published as L. C. Siebert, R. Mercuur, V. Dignum, J. van den Hoven, and C. Jonker. Improving Confidence in the Estimation of Values and Norms. In A. Aler Tubella, S. Cranefield, C. Frantz, F. Meneguzzi, and W. Vasconcelos, editors, *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIII*, pages 98–113, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72376-7.

or even impossible to write down precise rules describing them. One approach to deal with this variance is to treat aligning the actions of the AA with human values as a learning problem [132, 239]. However, estimating values and norms solely based on observed behavior may lead to ambiguous results. Different relative preferences towards values and norms can bring about the same behavior, i.e. there are many "reasons why" that might motivate a given observed behavior.

This chapter studies the conditions under which AAs are able to make confident estimates of values and norms from observed behavior. We studied the behavior of simulated agents driven by both values and social norms in the Ultimatum Game (UG), specifically in the proposer role. [199] has used the UG to study the relative importance of values across cultures. We focus on the relative importance of values and norms that guide individual decision making.

The main contribution of this chapter is twofold. First, we propose an agent-based model to play the UG, expanded from [174], which uses both values and norms for determining the actions of the agents. Second, we present a method for estimating an agent's relative preferences to a given set of values and, if necessary, to improve the confidence in these estimations. This improvement is realized by interacting with the proposer agent in the UG, either by a free exploration of the search space or by the use of counterfactuals.[1]

The remainder of the chapter is structured as follows. The following section presents a brief context about the UG and how we model values and norms in this context. Section 3 presents both the method to estimate the relative preference attributed to an agent's values and norm, and the methods to reduce ambiguity on the estimation. Section 4 presents the main results of this research and section 5 discusses these results. Finally, section 6 presents the conclusions.

5.2. Values and norms in the ultimatum game

5.2.1. Scenario: The Ultimatum Game

We simulate the behavior of the agents in the UG, first introduced by [115]. In the UG, two players negotiate over a fixed amount of money (the 'pie'). The proposer demands a portion of the pie, while the remainder is offered to the responder. The responder chooses to accept or reject this proposed split. If the responder chooses to 'accept', the proposed split is implemented. If the responder chooses to 'reject', both players get no money. The UG thus provides an environment where players have to make decisions about money based on their own judgment.

The remainder of the chapter uses a version of the UG with the following specifics:

- An experiment has 32 players: 16 proposers and 16 responders.

[1] The code for the simulation models, data analysis and ML algorithm is available online (see Appendix 8.3)

- One experiment has 20 rounds. One round in the UG comprises one demand for each proposer and one reply for each matched responder.
- The players are paired to a different player each round but do not change roles.
- Players are anonymous to each other.
- The pie size P is 1000[2].

These specifics are chosen for pragmatic reasons: the chapter uses an empirical dataset based on these specifics [48] and builds upon a model based on these specifics [174].

This chapter models the decision-making of humans in the UG by using values and norms, but there have been many different approaches to it since its first appearance in [115]. One reason for its popularity is that the UG is a classical example of where the canonical game-theoretical agent (i.e., the *homo economics* that only cares about maximizing its direct own welfare) falls short. The *homo economicus* only cares about its direct welfare and thus will accept any positive offer in the UG. Humans, in contrast, reject offers as high as 40% of the pie [203]. Behavioral economics has aimed to explain humans by incorporating learning [222], reputation [87] or other-regarding preferences. Recently, [174] used values and norms to explain UG behavior. Using a series of agent-based simulation experiments, [174] showed that the model produces aggregate behavior that falls within the 95% confidence interval wherein human behavior lies more often than the other tested agent models (e.g., a learning *homo economicus* model). Moreover, the model uses concepts that humans use in their explanations. Thus because the model has been shown to reproduce human behavior and uses explainable concepts, the model of [174] provides a starting point to estimate preference towards human values and norms for value alignment.

5.2.2. Simulated Human agent (SHA)

The simulated human agent (SHA) represents the human in the UG whose values and norms we aim to estimate. The model for the SHA is based on the value-based model and norm-based model presented in [174]. These models focus on the aggregate properties that emerge from pairing these agents in the UG. However, this chapter focuses on estimating relative preference towards values and norms of individual agents. Therefore, we are interested in an agent model where one agent uses *both* values and norms. The remainder of this section presents the SHA model in three parts: the value-based agent (from [174]), the norm-based agent (extension of [174]) and an agent that uses both values and norms (new).

Value-Based Agent

The value-based agent uses the importance an agent attributes to its values to determine what an agent (based on its values) should demand. [174] focuses on two

[2]For ease of presentation, we chose 1000 with no monetary unit to the pie size. Empirical work [203] shows that the effect of the pie size is relatively small.

values that are relevant in the UG: wealth and fairness and define di_a as the difference in importance an agent attributes to wealth versus fairness. Based on research by [229] and data on the UG [48], they then state a number of requirements for a function (see [174] for more details). First, they assume a (perfect) negative correlation between wealth and fairness. Second, valuing wealth is defined as wanting more money. Third, valuing fairness is defined as wanting an equal split of the pie. Fourth, agents should differ in how much money they demand, but not demand below 0.5P. The functions that map di_a to the demand $d \in \{0...P\} \subset \mathbb{Z}$ are given by

$$valueDemand(di_a) = \underset{d \in \{0...P\} \subset \mathbb{Z}}{\arg\max}\, u(d, di_a, P) \tag{5.1}$$

and

$$u(d, di_a, P) = -\frac{1.0 + 0.5di_a}{\frac{d}{P} + 0.5} - \frac{1.0 - 0.5di_a}{\frac{|0.5P - d|}{0.5P} + 0.5} \tag{5.2}$$

The agent thus calculates the utility for each demand d and returns the demand with the maximum utility. [174] shows these functions fulfill the stated requirements. By using (5.1) and (5.2), we enable our agent to choose a demand following theories on values [229] and [48] empirical results on the UG.

Norm-Based Agent

The norm-based agent uses the replies it observes from the other agent to determine what an agent (based on its norms) should demand. [174] follows theories by [53] and [90] and states that a norm exists for a particular person when that person perceives most other people do or expect it. [174] provides a translation from this definition to a model for the UG. In the UG, the proposer cannot see what other proposers do, because the proposer is only paired with responders. Therefore, the proposer uses the responses of the responder to form an idea of what is expected (i.e., the norm). [174] provides the following function to map the observed responses (OR_{ak}) seen by agent a up to the current round k to a demand $d \in \{0...P\} \subset \mathbb{Z}$,

$$normDemand(OR_{ak}) = \frac{\min\limits_{d \in RD_{ak}} d + \max\limits_{d \in AD_{ak}} d}{2} \tag{5.3}$$

where $RD_{ak} \subset OR_{ak}$ is the set of demands that the proposer a has seen rejected and $AD_{ak} \subset OR_{ak}$ is the set of demands that the proposer a has seen accepted. The function thus uses two indicators to estimate the norm: the minimum of the rejected demands (RD_{ak}) and the maximum of the accepted demands (AD_{ak}). The rejection of a given demand indicates that the demand is higher than expected, thus the norm is lower than the minimum rejected demand. The acceptance of a given demand indicates that the demand is lower than expected (or perfect), thus the norm is higher (or equal) to the accepted demand. Equation (5.3) determines the norm as the average of these two indicators. By using (5.3), we enable our agent to choose a demand following [53] and [90] theories on norms.

We extend the model provided by [174] to specify normDemand(OR_{ak}) for three edge cases: a case where there are no observed responses ($OR_{ak} = \emptyset$), a case with only observed rejections ($RD_{ak} = OR_{ak}$), and a case with only observed accepts ($AD_{ak} = OR_{ak}$):

- $OR_{ak} = \emptyset \rightarrow$ normDemand(OR_{ak}) is drawn from the normal distribution $N(561.8, 123.9)$, which is the normal distribution human demands follows[3].
- $RD_{ak} = OR_{ak} \rightarrow$ the agent averages between the minimum reject and half the pie $((\min_{d \in RD_{ak}} d + 0.5P)/2)$.
- $AD_{ak} = OR_{ak} \rightarrow$ the agent averages between the maximum accept and the full pie $((\max_{d \in AD_{ak}} d + P)/2)$.

In [174], the agent determines a randomized demand as the norm in all these cases. By specifying these edge cases, we improve the SHA model aiming to better represent human behavior.

5

Agent with Values and Norms

This chapter combines the value-based agent and the norm-based agent to model an agent that uses both values and norms. Values and norms are decision-making concepts the agent uses to decide what action is good and what action is bad [184]. [61] used values as more abstract ideals and norms as the more concrete context-dependent rules that follow from these ideals. Norms follow from values, but when agents copy the norms (but not values) from other agents the two can conflict. For example, one copies working over-hours although this is in conflict with it's own values. This chapter presents a model where an agent uses a weight (vw_a) to combine the best action based on values and the best action based on norms into the best action based on both values and norms.

For a proposer an action is defined as a demand $d \in \{0...P\} \subset \mathbb{Z}$. The demand d for an agent a in round k is determined by

$$d_{ak}(di_a, vw_a, OR_{ak}) = vw_a \times valueDemand(di_a) + (1 - vw_a) \times normDemand(OR_{ak}), \quad (5.4)$$

where di_a stands for the difference in importance an agent attributes to wealth versus the importance it attributes to fairness, $vw_a \in [0, 1]$ stands for the weight an agent attributes to its own values (versus the weight it attributes to norms) and OR_{ak} stands for the set of observed replies the agent has seen. Thus what an agent demands is the result of weighting the result of two functions described in [174]: the $valueDemand(di_a)$ (5.1) and the $normDemand(OR_{ak})$ (5.3).

The above focuses on the demands of the SHA-proposer (SHA-P), but to simulate the UG we also need SHA-responders (SHA-R). The SHA-R is defined the same

[3] The norm that is drawn from the normal distribution is not used as input for the norm in subsequent rounds (i.e., the agent does not memorize it).

way as to the proposer model except that it determines a threshold $t \in [0, P]$ (instead of a demand d) and has to base its norms on a set of observed demands (instead of observed responses). The SHA-R uses the same function as the SHA-P (5.4) to determine the threshold. If the demand is higher than the threshold the responder rejects it, if the demand is lower than or equal to the threshold accepts it. The responder determines a threshold appropriate for its values analogue to the proposer, that is, by using (5.1) and (5.2). The responder determines the norm for the threshold by averaging over the observed demands (instead of the observed responses). The proposed demand is thus seen as a signal from the proposer that this is what the proposer considers 'normal'.[4] By using a threshold t based on its values and the norm (from observed demands), the responder uses values and norms to reject or accept the proposed demand.

Above we presented a model that uses both values and norms using two agent-specific variables: the difference in importance (di) and the value-norm weight (vw). The difference in importance specifies how much an agent values wealth over fairness and is normally distributed over agents $N(\mu_{di}, \sigma_{di})$, being μ_{di} the average value and σ_{di} the standard distribution of a normal distribution for di. The value-norm weight specifies how much an agent weighs values over norms and is normally distributed over agents $N(\mu_{vw}, \sigma_{vw})$. Both variables are constant over the different rounds. However, the $normDemand(OR_{ak})$ differs per round as the observed replies vary. To reproduce the demands and responses humans give, the parameters μ_{di}, σ_{di}, μ_{vw} and σ_{vw} need to be calibrated to the right settings.

5.2.3. Calibrating the SHA to humans

We found that $\mu_{di} = 0.5$, $\sigma_{di} = 0.25$, $\mu_{vw} = -0.6$ and $\sigma_{vw} = 1.14$ produced simulated demands and responses that are closest to the empirical data on human demands and responses extracted from the dataset provided by [48]. This meta-study combines the data of 6 empirical studies to create a dataset of 5950 demands and replies made by humans in the same scenario as we describe in Section 5.2.1. We used five performance measures for which we compared the synthetic data on the SHA to the empirical data on real humans: average demand μ_d, standard deviation in demand σ_d, average acceptance rate μ_a, standard deviation in acceptance rate σ_a and the standard deviation in the demand solely based on values σ_{vd}.[5] The SHA is compared to humans on both round 1 and round 10. We used the following procedure to find the optimal parameter settings:

1. Run simulations of the UG in Repast Simphony with different parameter settings ($\mu_{di} \in [-1, 1]$ and $\mu_{vw} \in [0, 1]$) and different random seeds ($r \in [1, 30]$) and obtain the resulting demands and acceptance rate (μ_d, σ_d, μ_a, σ_a);

[4] To be exact, the proposed demand should be considered as what the proposer considers a 'normal' threshold. If it considered a higher threshold to be normal it would have demanded less, if it considered a lower threshold normal it would have demanded more.

[5] The standard deviation in the demand solely based on values σ_{vd} was added to ensure agents vary in what values they find important (di_a). The σ_{vd} for humans is postulated instead of extracted from empirical data.

Table 5.1: A comparison of the distribution of demands and responses between empirical data on humans and synthetic data on simulated human agents (SHAs) given the parameter settings with the best fit ($\mu_{di} = 0.5$, $\sigma_{di} = 0.25$, $\mu_{vw} = -0.6$ and $\sigma_{vw} = 1.14$).

		Performance Measures				
Round	Source	μ_d	σ_d	μ_a	σ_a	σ_{vd}
1	empirical (human)	561.8	128.9	0.806	0.40	128.9
	synthetic (SHA)	557.9	91.1	0.876	0.29	109.2
10	empirical (human)	584.2	98.66	0.868	0.34	122.5
	synthetic (SHA)	646.8	90.89	0.923	0.23	109.2

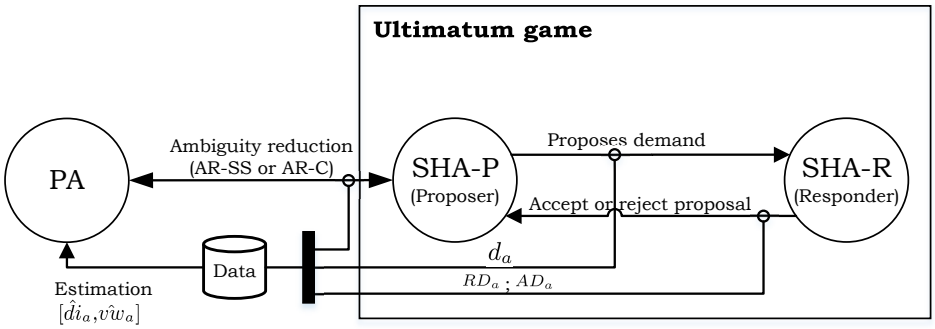


Figure 5.1: Conceptual model.

2. For each run: average the performance measures with the same parameter settings, but different random seeds;
3. For each parameter setting: calculate the normalized root mean square error (NRMSE) between the simulated results and the human results;
4. Pick the parameter setting with a minimal NRMSE.

Table 5.1 compares the resulting demands and responses for the SHA (when NRMSE is minimized) and the empirical data on human results. We interpret these results as that our SHA can produce a distribution of demands and accepts that it is fairly close to that of humans (NRMSE = 11.0). The remainder of the chapter uses these parameter settings to simulate human behavior based on values and norms.

5.3. Estimating relative preferences on values and norms

The conceptual model of the estimation process is presented in Fig. 5.1. The *Profiling Agent* (PA) is responsible for estimating the relative preference of a given SHA-P towards values and norms ($[\hat{di}_a, \hat{vw}_a]$), and whenever necessary, interact with the SHA-P to reduce ambiguity in the estimation.

The problem of estimating values and norms in the context of this work is translated to the estimation of di_a and vw_a, represented as $[\hat{di}_a, \hat{vw}_a]$. The estimation of relative preferences from behavior was assessed in different ways, such as via Inverse Reinforcement Learning (IRL) algorithms [1, 116, 158, 185] and learning a utility function with influence diagrams [194]. In this work, we use the UG as a simple context (two values and one norm), aiming for clarity on understanding the conditions necessary to properly estimate the relative value and norm preference from behavior and, whenever necessary, how to reduce ambiguity.

In the context of the UG we can perform the estimation of the relative preferences by a exhaustive search process on the estimators, according to:

$$[\hat{di}_a, \hat{vw}_a] = \underset{di_a \in [-0.15;1.79]; vw_a \in [0;1]}{\arg\min} \frac{\sum_{k=1}^{m} \left| \hat{d}_{ak}(di_a, vw_a, OR_{ak}) - d_{ak} \right|}{k} \quad (5.5)$$

where $\hat{d_{ak}}$ is calculated by (5.4), and m represents the number of rounds from the UG analyzed in the estimation. In short, the estimators $\hat{di}_a$ and $\hat{vw}_a$ are defined as the preference set that minimizes the deviation between the observed demand (d_{ak}) and the demand that was calculated by an exhaustive search process ($\hat{d_{ak}}$). The range for di_a which has an effect on the demand by the SHA-P is $[-0.15; 1.79]$. With $vw_a \in [0; 1]$, and using a step of 0.01 to explore both variables, 19,392 evaluations of the fitness function are necessary for each estimation process.

5.3.1. Reducing ambiguity

Ambiguity arises whenever the number of elements of $\hat{di}_a$ and $\hat{vw}_a$ is greater than 1. The PA will try to reduce this ambiguity by taking the role of the *responder* on the UG. This interaction might be interpreted as an elicitation process or, in simple, the PA will ask the SHA-P "questions". Different ways of "asking these questions" may produce higher or lower quality answers, and consequently a higher or lower change in the confidence on the estimation. We will explore two different approaches to how the PA interacts with the SHA-P. These interactions are a "side-game", i.e. changes in OR_a will not, by definition, influence future actions of the SHA-P.

Ambiguity Reduction by exploring the Search Space (AR-SS)

The approach to ambiguity reduction via exploring the search space ($AR-SS$) poses the following "question" to the SHA: *What would your next demand be?*. Based on the demand ("answer") provided by the SHA-P, the PA will reject or accept it to explore the search space (Algorithm 1). This approach increases the search space, i.e. the distribution of OR_a in the dataset, by rejecting demands lower than rejected before and accepting demands higher than accepted before.

Algorithm 1: Ambiguity Reduction by exploring the Search Space ($AR-SS$)

input : $\hat{di}_a$, $v\hat{w}_a$, max_{int},RD_a,AD_a,OR_a, d_a, k
output: $\hat{di}_a$, $v\hat{w}_a$,RD_a,AD_a,OR_a, $count$

$n_{solutions} \leftarrow$number of solutions $([\hat{di}_a, v\hat{w}_a])$
$k \leftarrow k+1$
Calculate OR_{ak}
$count \leftarrow 0$
while $n_{solutions} > 1$ *AND* $count < max_{int}$ **do**
  Calculate $\hat{d}_{ak}(\hat{di}_a, v\hat{w}_a, OR_{ak})$ (5.4)
  Include OR_{ak} and $\hat{d}_{ak}$ to the data set
  Estimate $[\hat{di}_a, v\hat{w}_a]$ (5.5)
  $n_{solutions} \leftarrow$number of solutions $([\hat{di}_a, v\hat{w}_a])$
  if $d_{ak} < min(RD_a)$ *OR* $\nexists\ RD_a$ **then**
    $RD_{ak} \leftarrow d_{ak}$
  else if $d_{ak} > max(AD_a)$ *OR* $\nexists\ AD_a$ **then**
    $AD_{ak} \leftarrow d_{ak}$
  $k \leftarrow k+1$
  Calculate $OR_{a,k}$
  $count \leftarrow count+1$

Ambiguity Reduction via Counterfactuals (AR-C)

Counterfactuals are mental representations of alternatives to events that have already occurred [218], frequently represented by conditional propositions related to questions in the "what if" form. Philosophical discussion on counterfactuals has been present for ages, including works from David Hume and John Stuart Mill [205], and it is considered an intrinsic element of causality. People use counterfactuals often in daily life to create alternatives to reality guided by rational principles.

Our approach to reduce ambiguity on the estimation of values and norms via counterfactual ($AR-C$) poses the following "question" to the SHA-P: *What would your next demand be if your opponent had accepted instead of rejected your proposal on round "x"?* or *What would your next demand be if your opponent had rejected instead of accepting your proposal on round "x"?*. As presented in Algorithm 2, the PA will 'ask' the 'question' related to the round that will lead to a broader search space in terms of the observed social norm (OR_a).

5.4. Results

To test the methods presented in this chapter we analyzed the behavior of agents acting as proponents (SHA-P) in 100 runs of the UG (each with 20 rounds), in terms of precision and confidence. The first will be evaluated by the root mean squared error (RMSE) between the observed demand (d) and the estimated demand

Algorithm 2: Ambiguity Reduction via Counterfactuals (AR-C)

input : $\hat{di}_a$, $\hat{vw}_a$, RD_a,AD_a,OR_a, d_a,k
output: $\hat{di}_a$, $\hat{vw}_a$,RD_a,AD_a,OR_a, $count$

$max_{int} \leftarrow k$
$count \leftarrow 0$
$n_{solutions} \leftarrow$number of solutions $([\hat{di}_a, \hat{vw}_a])$
while $n_{solutions} > 1$ *AND* $count < max_{int}$ **do**
  for $i = 1 : k$ **do**
    if $\exists AD_{ai}$ *(Proposal was **accepted** in round i)* **then**
      if $d_{ai} < min(RD_a)$ *OR* $\nexists\ RD_a$ **then**
        $RD_{ai} \leftarrow d_{ai}$
    else if $\exists RD_{ai}$ *(Proposal was **rejected** in round i)* **then**
      if $d_{ai} > max(AD_a)$ *OR* $\nexists\ AD_a$ **then**
        $AD_{ai} \leftarrow d_{ai}$
    Calculate OR_{ai}
    $score_i \leftarrow min(|OR_{ai} - OR_a|)$
  $k \leftarrow k + 1$
  $OR_{ak} \leftarrow OR_{az}$, where z is the iteration where $score$ is maximum
  Calculate $\hat{d}_{ak}(\hat{di}_a, \hat{vw}_a, OR_{ak})$ (5.4)
  Include OR_{ak} and d_{new} to the data set
  Estimate $[\hat{di}_a, \hat{vw}_a]$ (5.5)
  $n_{solutions} \leftarrow$number of solutions $([\hat{di}_a, \hat{vw}_a])$
  $count \leftarrow count + 1$

($\hat{d}$), while the latter by the number of elements in $[\hat{di}_a, \hat{vw}_a]$. A small number of elements represent high confidence (low ambiguity), while a large number of solutions represent low confidence (high ambiguity). It is guaranteed the number of elements of $[\hat{di}_a, \hat{vw}_a]$ is greater than zero, since we perform an exhaustive search in the complete range of di_a and vw_a. With these results, we aim to analyze the conditions under which the proposed methods can estimate preferences on values and norms in a precise and confident manner.

5.4.1. Estimation of values and norms

The precision in the estimation increased with the number of rounds used, given that an estimation is made by observing at least four rounds (Fig. 5.2. (a)). Using less than four rounds the estimated values and norms ($[\hat{di}_a, \hat{vw}_a]$) predicted a demand ($\hat{d}$) that is very close to the real demand[6], but most of these estimations will not be able to generalize among other contexts (i.e. other observed response

[6]Given the deterministic model of the SHA, it might be expected that the RMSE should tend to zero. This is not the case because $[d, valueDemand, normDemand] \in \mathbb{Z}$, and therefore a rounding operator is used.

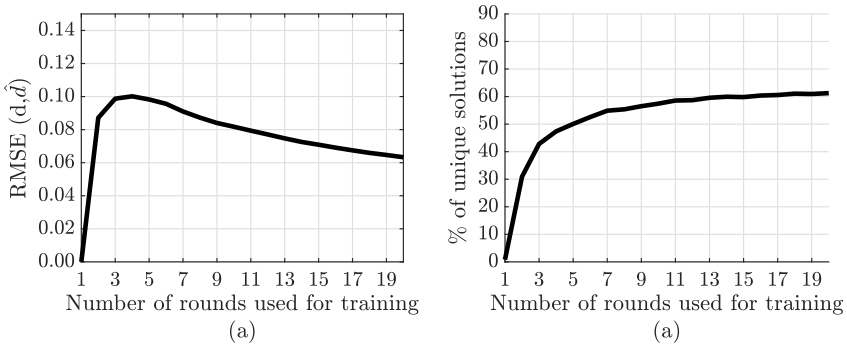


Figure 5.2: (a) Precision of the estimation process. (b) Percentage of unique solutions for the estimation process.

- OR_a).

When considering between one and four rounds the estimations were ambiguous: on average 69.2 different sets of estimations led to the same demand (Fig. 5.2. (b)). Nevertheless, the confidence in the estimation increased with the number of rounds used. The steep curve reaching round four appears in both figures, showing a correlation between this initial precision in the estimation of the demand with the ambiguity on estimating preferences toward values and norms. Even with an increasing number of rounds observed, the percentual of unique solutions reached an average of only ca. 60%. These different estimations may lead to undesired properties of an agent that wants to act on behalf of the true values and norms of a given person.

5

5.4.2. Reducing ambiguity

Both proposed methods were able to increase confidence in the estimation of preferences on values and norms ($[\hat{di}_a, \hat{vw}_a]$). Comparing Fig. 5.3. (a) with Fig. 5.2. (a), the RMSE between the calculated and observed demand are slightly higher, but still may be considered adequate in absolute values, given that $d \in \{0...1000\} \subset \mathbb{Z}$.

The methods $AR - SS$ and $AR - C$ did not increase confidence in the same manner, differing in their applicability. While $AR - SS$ was able to reach estimations with less ambiguity than $AR - C$ (Fig. 5.3. (b)), it required in average of 4.9 more interactions with the user (Fig. 5.3. (c))[7]. In 33.2% of the cases where the estimation was ambiguous, $AR - SS$ could provide a unique solution, while $AR - C$ provided it for 27.2% of the cases. Considering both the confidence and the number of interactions needed, $AR - C$ increased the number of unique solutions by 16.5% per interaction with the SHA-P, while $AR - SS$ increased it by 4.1% per interaction.

Table 5.2 summarizes the results when the estimation is performed using 10 rounds of observed data. We can see that precision varied only slightly, but the

[7]The x-axis in Fig. 5.3 relates to the number of rounds used during the initial estimation process, described in section 5.4.1. The additional number of interactions performed by each method to improve the confidence in the estimations is not included in the x-axis but presented in Fig. 5.3. (c).

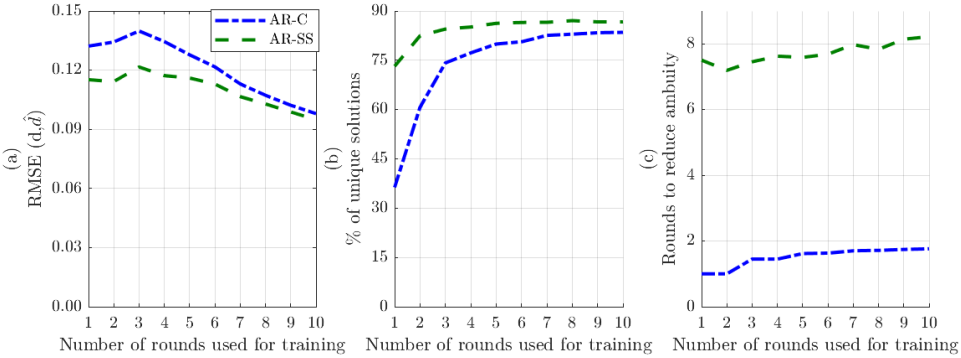


Figure 5.3: Reducing ambiguity: (a) Precision; (b) Unique solutions; (c) Number of interactions made for reducing ambiguity.

Table 5.2: Summary of the results for the estimation using 10 rounds, and for the subsequent ambiguity reduction methods.

				Standard deviation		
Method	Precision	% Unique	Rounds on AR	OR_a	$\hat{di}_a$	$\hat{vw}_a$
Estimation	0.082	57.4	-	36.1	0.021	0.008
AR-SS	0.095	86.7	8.2	42.6	0.011	0.003
AR-C	0.098	83.5	1.8	45.5	0.013	0.003

confidence in the estimation increased significantly when using any of the *ambiguity reduction* methods proposed. The columns related to the "Standard deviation" on Table 5.2 provide parameters to understand the dispersion of the search space[8] and of the estimations ($[\hat{di}_a, \hat{vw}_a]$). While the search space scatters with the ambiguity reduction methods, there is a reduction in the dispersion for the estimations. In other words, as the search space covered by OR_a increases, estimations become less ambiguous (and more gathered).

5.5. Discussion

AAs must be able to estimate a human's relative preference towards values and norms to align their behavior with them. This is not an easy task since a different set of preferences might lead to the same observed behavior (ambiguity problem). [185] demonstrated that 'normative assumptions' are needed to estimate an agent's reward function (in our case, values and norms). In this work, we proposed an ABM that uses both values and norms to account for these 'normative assumptions', and methods to estimate the SHA preferences and reduce the ambiguity in these estimations.

The first contribution, the proposed ABM, was built on previous works [161, 184], and especially [174]. We extended the ABM works to use *both* values and

[8] in our case OR_a: given that preference to values and norms are constant, demand is defined according to 5.4)

norms for determining the actions of the agents. The model was calibrated to represent human behavior from empirical data. Future works can improve this model by specifying in more detail the use of norms and values, considering other values than wealth and fairness, constraining behavior by using deontic operators [174], and incorporating models that represent human bounded rationality.

The second contribution of this work was focused on understanding to what extent the proposed methods were able to estimate the relative importance attributed to values and norms by a given agent. We show that even when considering 'normative' assumptions, represented by the PA's knowledge of the decision-making process of the SHA in a relatively simple context (the UG), ambiguity may still be present on estimating preferences for value alignment. To ignore this ambiguity might lead to great regret and unethical actions.

In the experiment performed we observed two general conditions under which an estimation of the preferences was possible, namely: heterogeneity on the observations and a sufficient number of observations. When observing at least four rounds of the UG, only 50 to 60% of the estimations were unambiguous. We proposed two methods for reducing ambiguity: one via exploring the search space ($AR-SS$) and one using counterfactuals ($AR-C$). The first approach, $AR-SS$, interacts with the SHA-P considering the last observed norm (OR_{ak}), which might be suitable for interacting with people in the real world due to the influence of short-term memory on decision making [63]. The latter approach, $AR-C$, uses counterfactuals, which are a part of causal reasoning. The results presented in the previous section show that targeting "imaginative" scenarios related to a specific round of the UG significantly increase the efficiency of the process. Nevertheless, it might be very demanding for a person to be able to go through this thought imaginary process and remind the precise social norm at a specific point in time. We can conclude that the $AR-SS$ method is more suitable where there are no restrictions regarding the number of interactions with the user, while $AR-C$ can improve confidence in situations where a limited number of interactions is desired.

Both methods act with the assumption that the only way the PA can interact with the SHA is by taking the role of the responder in the UG. Going beyond this assumption, we also evaluated a different approach: the PA can directly define a value for the social norm (OR_{ak}). If we consider $OR_a \in [0;1000]$, the ambiguity was almost eliminated: 97% of unique solutions on average, considering up to 20 interactions. If we consider a more realistic assumption $OR_{ak} \in [500;1000]$ the results were, in terms of the percentage of unique solutions, between the levels found by $AR-SS$ and $AR-C$ when using less than 6 rounds for training, and slightly better than $AR-SS$ when using 6 or more rounds. Future works can evaluate this hypothesis. We suggest that such experiments be done either considering improved models of human memory and cognition process or in laboratory settings.

The limitations of this work include the impossibility of directly generalizing the findings and methods to other contexts, the assumption that values are stable, and the lack of testing of the approach with humans in realistic settings as well as in more complex settings.

5.6. Conclusion

This chapter aimed to investigate to what extent an AA might be able to estimate the relative preferences attribute to human values and norms, including methods to reduce ambiguity. Insight into the use of models to support the estimation of values and norms was obtained during the discussions, mainly the need for heterogeneity on the observations and also ways to reduce ambiguity. Especially the use of counterfactuals via the $AR - C$ approach showed it can be of great value in terms of a trade-off between increasing efficiency and avoiding excessive interactions/questions with humans. We showed that even in a simple context, considering models that represent values and norms ('normative assumptions'), and using a exhaustive search process, ambiguity cannot be easily be avoided in estimation of preference on values and norms. To ignore this ambiguity might lead to great regret and misalignment between machine behavior and human values and norms.

Acknowledgements

This work was supported by the AiTech initiative of the Delft University of Technology.

6

Identifying Social Bottlenecks in Hospitals: Modelling the Social Practices of an Emergency Room

You see, in many of the organizations I've worked with, people can't talk about their feelings. Nobody cares about what you feel and need. It's all about production. But when you don't express your feelings and needs, when you just keep going into intellectual discussions, you end up using your time unproductively by not getting down to the root of the problem.

Speak Peace in a World of Conflict, Marshall B. Rosenberg

This chapter has been published as R. Mercuur, J. B. Larsen, and V. Dignum, "Modelling the social practices of an emergency room to ensure staff and patient wellbeing," Belgian/Netherlands Artif. Intell. Conf., pp. 133–147, 2018 [173].

6.1. Introduction

A better understanding of the impact of activities in an emergency room (ER) on patients and staff would improve the wellbeing for both of them [162]. A decision support tool that has knowledge about ongoing activities could assist ER management in directing interventions by identifying what activities take place, what causes them to take place and who are typically involved in the activities. An ER is a complex social multi-agent system, where staff members should understand the needs of patients, what their colleagues expect of them and how the treatment normally goes about [241]. Focussing on this social dimension in these decision support tools could increase their realism [196] and therefore their potential in giving helpful insights into the domain.

So far ER models have mainly focussed on the patient flow, but not the interaction of the staff. An ER is modelled from a control flow perspective based on explicit regulations and clinical guidelines [219]. This is useful to identify possible process bottlenecks, but offers little insight into their social causes. A possible cause for the lack of sociality in decision support tools for ER might be that there is no formal model that expresses this social dimension.

To use a formal reasoner to identify social bottlenecks three steps are involved (1) fairly representing the social dimension of an ER in a model (2) formalizing the model and using a formal reasoner to check if the model satisfies certain properties (3) translating these results back to interventions in the real world ER. This chapter focusses on the second step by defining a precise, unambiguous and consistent semantics to support decision support tools. In particular, we aim to give a basis for formal reasoners that can infer helpful new social knowledge about an ER. Building upon [77, 214, 235], we use social practices to capture the social dimension of a system. The social practices in a system are the routinised actions that are (to some extent) similar for all agents such as, diagnosing or treating the patient. We use the Web Ontology Language (OWL) [28, 118, 129, 261], based on Description Logic, and the Protégé tool to formalize a social practice meta-model and apply it to the ER domain. This results in a formal ontology that includes a social practice meta-model and a domain specific ER model that can be used for decision support tools. We demonstrate, with data based on a visit to Herlev Hospital in Denmark, that we can express and verify whether a number of social properties that one could desire of an ER holds true in our model. We thus provide a descriptive model of an ER and not a prescriptive model of how an ER should be run. This serves as a proof of concept to show that our formalization can be used to infer helpful new knowledge about a system. These results serve not only as a first step to better decision support tools for ER, but also serves as an example for formalizing the social dimension of other multi-agent systems.

Section 2 presents some background knowledge on social practices. Section 6.3 introduces the social dimension of the ER domain, its empirical grounding, and lists a number of exemplary social properties one could desire of an ER. Section 6.4 formalizes the social practice meta-model and our ER use case. Section 6.5 formalizes the earlier stated social properties and uses a formal reasoner to verify them (i.e., infer new knowledge about the system.) We end the chapter by showing how our

formalization is unique in this ability to capture social properties by comparing it to other agent meta-models.

6.2. Background

The concept of social practices stems from sociology, and aims to depict people's 'doings and sayings' [227, p. 86], such as dining, commuting and greeting. [235] recently revived the concept for its ability to highlight that our actions can be captured in routines that are similar for many people. For example, doctors and nurses follow similar routines in treating a patient. [77] proposed to use social practices to capture the social dimension of agent decision-making. They connected the concept to more standard agent concepts, such as actions, plans and norms. This section introduces a meta-model for a social practice agent that is expressed in the Unified Modelling Language (UML). The remainder of this chapter focuses on the formalization of this UML-diagram in OWL. This formalization provides an unambiguous basis for decision support tools in social systems such as the ER. Future work will focus on how the UML-diagram is grounded in the social practice literature and the justification of the model choices.

Figure 6.1 shows the social practice meta-model in a UML-diagram [223]. The main classes for a social practice model are activities (e.g., diagnosing, assign team), agents (e.g., nurses, doctors), competences (e.g., perform triage, managing team), context elements (e.g., phone, bed, triage room, other nurses) and values. Values here refer to human values, that is, 'what one finds important in life' [209], such as health or education. The social practice is an interconnection of (1) activities and (2) related associations as depicted by the grey box in Figure 6.1. For example, the social practice of acute treatment consists of several activities, such as assigning teams or performing triage. Figure 6.2 shows all the activities that the social practice of acute treatment comprises. Section 6.3 will further explain what these mean, for now it is important to understand the structure of this activity tree. Shallow nodes are more abstract activities (e.g., assign to team) and deeper nodes are more concrete ones (e.g., inform patient).[1] The type of an activity (respectively, `AbstractAction` and `Action`) is captured in the `type` attribute in the UML. The social practice connects these different activities with the `Implementation` association. If activity A implements activity B this means that A is a way of or a part of doing B.

Implementation is the first of several activity-associations specified in Table 6.1. Most associations are fairly self-explanatory, however the `Trigger` and `Strategy` association are a bit more complex. Following [267], triggers are the basis for habitual behaviour. If an agent is near a context element that has a trigger association with an activity, then it will do that activity automatically (without for example considering its values). Following [53], strategies are the basis for expectation about the actions of others. If an agent believes that (the personal part of) activity A is a strategy for activity B, then it believes that other agents usually

[1] Note that what is considers a concrete activity is a modelling choice that differs per study. The choices made here are a proof of concept that need to be extended based on the purpose of the study and domain-specific empirical data.

implement activity B by doing activity A. Fundamental to social practices are that many of these activity associations are similar (or: shared) for most agents. Section 6.4.2 will explain how we capture this similarity in the model.

Table 6.1: The associations attached to an activity and their specification.

Association	Specification
Implementation	which activities are a way of or a part of doing the activity
Affordance	which context elements are needed to do the activity
RequiredCompetence	which competences are needed to do the activity
Belief	which activities an agent has beliefs about
RelatedValue	which values are promoted or demoted by the activity
Trigger	which context elements habitually start the activity
Strategy	which activities usually implement the activity

The agent furthermore has two associations, which plays a role in choosing the activities it will do: `HasCompetence` and `HasValue`. The `HasCompetence` association links possible skills to the agent who masters those. The `HasValue` association captures if an agent finds that value important.

6.3. ER Use Case

6.3.1. ER Description

An ER acts as the entry point for most hospitals and a wide variety of patients arrive there on short notice. The purpose of an ER is to provide immediate treatment for the patient and identify the further course of action. The general process of patient treatment can be seen as consisting of the following phases:

1. Contact ER: The secretary registers the patient who contacts the ER department.
2. Triage: A nurse performs triage on the patient in order to identify how urgent or life critical the patient is.
3. Tests: Doctors and nurses perform tests on the patients.
4. Diagnosis: Doctors give a (partial) diagnosis and perform treatment.
5. Plan: Doctors, together with the patient, establish a plan for further course of action.

Staff members thus comprise doctors, nurses, but also persons who are in charge of administrative tasks and of being in contact with patients. In some cases, an ER also serves as a learning facility for healthcare staff in training. It is common to have trainees participate in the treatment to get work experience as part of their education. Some trainees may have enough experience to carry out tasks by themselves, while others must be accompanied by more experienced staff.

The staff thus needs to cooperate and coordinate their actions, while being flexible enough to handle a wide range of patients. This means that staff members should understand the needs of other staff members and have an idea of what

guides their actions. A staff member should be aware of the culture of the organization by understanding what is important and how things normally go about. He or she needs to know what is expected of him and what not. All of this facilitates the teamwork that is needed for a properly running an ER. As should be evident from the above description, social interaction is an important part of an ER, which makes it an interesting domain to apply social practices.

6.3.2. Empirical Grounding

The ER description is partially based on observations from a half-day tour at the ER department at Herlev Hospital in Denmark. The tour was led by head nurses from the department and consisted of a visit to the main reception, the trauma reception, and the areas of the three specialist teams of the ER department. During the tour the head nurses explained the rooms in the department, the organization, general work procedures, and how a typical work day went like. After the tour we observed some of the staff members in the specialist teams for a few hours, asking a few questions whenever possible. During the visit several staff members relayed some personal views on their work, including what they considered important in doing it. Note that to our knowledge there are no empirical studies of the social practices in an ER (e.g., [6, 23, 264] only provide high-level analytical statements), therefore our work is based on observations from this half-day tour. To go beyond the proof of concept presented in this chapter a more extensive empirical study is needed.

Summary of relevant observations

The general attitude in the ER department is geared towards being flexible and accommodating the needs of the patients. The established patient treatment procedures and task distributions are suitable for the treatment of most patients. However, the staff is open towards making adjustments if they believe the adjustments can help. The following list comprises some examples of social intelligence we observed:

- Sending samples to analysis is supposed to be done by nurses, but because many patients required attention, the head nurse helped with sending samples to analysis to give the nurses more time to attending patients.
- The secretary is supposed to receive patient transports but the head nurse helped out with this as well sometimes, leaving a note that the secretary would later then register in the system.
- In the middle of the staff room there was a box with current tasks. Nurse students carried out some of the tasks under supervision of nurses who assigned the tasks to the students. When the number of tasks had grown large, the head nurse called one of the other head nurses to ask if they could help with some of the tasks.
- The secretary assisted with attending patients to the extent possible.

6.3.3. Desirable Social Properties

To evaluate if we can indeed express and reason about the social dimension of the ER, we state a number of concrete social properties one could desire, which reflects some of the observations we made during the tour:

1. The staff understand the needs of the patients.
2. A head nurse can cover some of the necessary tasks of the secretary.
3. The staff can help each other out, because they know the equipment the others need.
4. Nurses should follow directions from the head nurse.
5. The length of stay of patients is not longer than needed.

To give an idea of the scope to which our proof-of-concept can be generalized we aimed at a diverse set that relates to both patient and staff, both ambiguous and unambiguous and general and domain-specific statements. To make claims about how well our model could capture any social property lies outside the scope of this chapter. Section 6.5 shows that we can express the first three complex social statements with our formalization and verify their truth using a formal reasoner.

6.3.4. Social Building Blocks

The core of the ER model will be the activity tree that consists of instantiations of the `Activity` class and the `Implementation` relation between them (see Figure 6.2). The activity tree roughly shows the phases of acute treatment represented as social practice activities. The root of the tree represents the top action, the leafs represent the actions and the nodes in between represent abstract actions. It breaks the Acute Treatment activity down into more concrete activities to match the phases of patient treatment in ER. Note that the activity tree includes administrative activities such as assigning a patient to a specialized team within ER. We have included administrative activities to show how social practices can provide insight into these activities as well as activities that involve the patient.

The ontology is based on this activity tree, but also provides more details of social practices in line with the UML of Figure 6.1. Table 6.2 provides an overview of agents, values and competences in the ER case. We only consider a small number of values and competences from the use case in order to make it clear how they are represented and reasoned with.

6.4. Formalization

6.4.1. Meta Model

This subsection explains how the social practice meta-model described in Section 6.2 is formally captured in the Protégé tool. This comprises (1) translating UML concepts into equivalent assertions in the OWL syntax, (2) solving ambiguity and (3) making additional assertions that could not be captured in the UML syntax. The full formalization expressed in OWL can be found on GitHub (see Appendix 8.3).

Table 6.2: Overview of agents, the values they adhere to and their competences.

Agent	Values	Competences
Patient	Health	Feedback
Secretary	Prof.	IT
Head nurse	Prof., Educ.	IT, team
Nurse	Prof.	Triage, social
Exp. Trainee	Educ.	Triage, social
New Trainee	Educ.	Social
Doctor	Prof.	Medical

Class Hierarchy

The UML classes can be translated one to one to OWL classes. However, we also need to specify the relation between classes. We can specify that an individual can not be a member of two different classes with the disjoint class axiom. Note that the `Agent` class and the `ContextElement` class are not disjoint. The UML shows this with the generalization arrow that goes from `Agent` to `ContextElement`. Following [22] such an association can be captured in OWL with the subclass axiom.[2] In addition, we choose to capture the different activity types (i.e., `Action`, `AbstractAction` and `TopAction`) and different context elements (i.e., `Agent`, `Resource`, `Place`) as separate classes instead of attributes. This choice fits well with the notion of classes and avoids introducing an auxiliary typing scheme just for activities. This thus introduces three new disjoint classes (`Action`, `AbstractAction` and `TopAction`), which are all subclasses of the `Activity` class. This class correspondence can be captured in the 'disjoint union of' axiom.

Class Associations

Each association is translated to an object property in OWL (except for the earlier discussed 'generalization'). Object properties are defined by which two classes they connect (called respectively domain and range). We want to highlight a few interesting modelling choices. First, given the open world assumption of OWL[3], we do not need to explicitly assert statements in OWL about the possibility of any number of associations (like is done in UML with the '0...*' symbol). However, the '1' multiplicity, that restricts the amount of classes does need to be asserted. We simply capture this in an object property cardinality restriction. These restrictions are not expressed as isolated statements, but instead as a restriction on a class. For example, we restrict the `SharedPart` class by saying it is a subclass of all entities that implement exactly one activity. Second, some of the associations in the UML have attributes. Given that OWL supports the notion of a subproperty, we capture these attributes in the 'disjoint union of' axiom (analogue to the `type` attribute of the `Activity` class).

[2] For present purposes this suffices but for other OWL extensions some care should be taken in distinguishing subclasses, subtypes and the is-a relation.

[3] The OWL reasoner assumes that as long as we do not explicitly state something *not* to be the case, it is possible.

Complex Rules

Using SWRL[4] we can assert a number of propositions that correspond to the semantics of social practices, but could not be expressed in UML. Not only are these SWRL rules crucial to making inferences about social properties, they also depict new insights we gained due to formalizing the UML in the OWL language. First, it allows us to assert several propositions regarding the relation of `Action`, `AbstractAction` and `TopAction`. Top actions are defined as activities that implement nothing, whereas actions are defined as activities that are implemented by nothing. Lastly, abstract actions can be found in the middle of an activity tree, implementing some other activities and being implemented by some other activities. Second, we can sometimes infer that if an object property relates to one activity (e.g., `affordance`, `relatedValue`, `requiredCompetence`), then it should also relate to another activity. For example, if `TriageExperience` is a `requiredCompetence` for `Triage`, then it should clearly also be a `requiredCompetence` for the `AssignStaff` activity, as this is a part of `Triage`. In other words, given that `AssignStaff` implements `Triage` it inherits some of its object properties. Not only the competences of a parent activity are inherited, but also its related values and affordances. For example, if the context element `ER-room` affords `AcuteTreatment` it also affords `Triage`. Third, as mentioned a strategy captures which activities usually implement an activity. To be more exact, a strategy connects the activity A and activity B, if an agent believes that activity A is usually done as an implementation of B. Thus for a strategy to exist between A and B, A at least has to be an implementation of B.

6.4.2. Model

Having a formal meta-model for social practices, we use it to formalize a model for the ER use case. The model asserts facts that are roughly based on our half-day tour at Herlev Hospital in Denmark. In a later section, we use formal reasoning to verify properties that follow from these assumptions. Note that to go beyond the proof of concept presented in this chapter a more extensive empirical study is needed.

Shared and Personal Activities

Social practices capture the social world in terms of shared activity associations. This means agents can have the same beliefs about activities as other agents. In our model, for example, all agents believe that the activity `Triage` requires the `Triage` competence. However, there are also personal views on activities. For example, a patient might believe `Triage` only relates to the value of `Health`, while a trainee thinks it also relates to the value of `Education`. This is reflected in the model by having multiple instantiations of an activity: some shared and some personal. For example, there is one instantiation called `SharedTriage`, which all agents believe. Shared associations such as that the `Triage` activity requires the `Triage` competence are associated to this instantiation. There is also an instantiation called `TraineeTriage`, which is only believed by trainees. The

[4] SWRL is a language for extending an OWL ontology with Horn-like rules [129].

personal view that the value of Education is associated with Triage is attached to this instantiation. The fact that agents can assume some beliefs are similar for other agents is what makes it that we can infer *social* knowledge (as we will show in Section 6.5) and is what makes our ontology a unique tool (as we will discuss in Section 6.6).

Activity Tree

The parts of the meta-model that constitute the activity tree are the classes Activity, TopAction, AbstractAction, Action, and SharedPart, and the object properties implementationPartOf and implementationAllOf. We translate each box in the activity tree into the corresponding activity so that the root is an individual of TopAction, the leaves are individuals of Action, and all other boxes are individuals of AbstractAction. Initially we create one shared activity for each box, and later we create personal activities for the agents.

Agents and Strategies

The agents represent persons in an ER such as a nurse or a patient. For each agent we define a named individual of the Agent class. We assert that these agents know about an activity using the belief object property. In addition, we use the strategy object property to assert which activities agents believe as common ways of doing things.

Competences, Values and Context Elements

For each activity, we assert agent competences that are required for that activity, and which agents own those competences. For example, we assert that it is commonly believed that triage requires triaging competences by asserting Triage as a required competence of SharedTriage. We then assert that a nurse has the triage competence by using the HasCompetence object property. We also assert values that the agents adhere to and associate with the different activities. For example, we assert that it is commonly believed that triage promotes health, by asserting Health as a promoted value of Triage. Finally, we assert context elements that affords activities taking place, and which can trigger habits of some of the agents. For example, we assert that the ER facility affords contacting ER by asserting that ER is an affordance of SharedContact_ER and that taking contact in the ER is a habit of the patient by asserting it a trigger of PatientContact_ER.

6.5. Formalization & Evaluation of Properties

This section first formalizes the properties introduced in Section 6.3.3. The formalization follows three steps (1) expressing the statement in the more precise social practice terminology, (2) rephrasing it into a query and (3) rewriting it in the formal SPARQL syntax. We need to rephrase the properties into queries as Protégé can not evaluate the truth of a proposition, but rather gives back the lists of individuals that satisfy the query. In addition, this sections aims to verify the truth of the propositions, by evaluating what the query returns. We aim to demonstrate that our formalization allows us to reason about knowledge we did not assert. We leave claims about actual specific ER departments for future work.

Property 1 We make the property more precise by specifying what it means to 'need' and what it means to 'understand'. We specify the 'needs' of a patient as those context elements and competences that are required for (or afford) those activities the patient finds important. Important activities are those that promote values the patient adheres to. We translate 'understanding' something to having beliefs about something. In this case, the staff thus needs to have beliefs about certain competences and context elements that are important to the patient. Note that if in our formalism an agent have beliefs about an activity, this implies it believes the related competences and affordances. We can thus specify this property as 'All the activities that a patient believes promote values that are adhered to by the (same) patient, the staff have beliefs about'. We can query which agents have beliefs about the values that are important to the patient and check to what extent the staff members are part of this list.

```
SELECT ?agent ?value ?activity
WHERE {
    ?agent a:belief ?activity.
    a:Patient a:HasValue ?value.
    ?activity a:promotedValue ?value}
```

The formal reasoner returns a list of important activities and agents that have beliefs about these activities. In our proof of concept, we find that for every activity important to the patient, there is at least one staff member that has beliefs about this activity. However, not *all* the staff members have beliefs about the needs of the patient. For example, contacting the ER department is important for the patient, because it promotes the patient's health. One of the needs of the patient is thus the context element 'phone' that affords contacting the ER department. However, the formal reasoner shows that in our use case only the overall head nurse is aware of this.

Property 2 We make the property more precise by specifying what 'can cover' ,'necessary tasks' and 'tasks of the secretary' mean. We specify the 'tasks of the secretary' as those tasks the secretary can be triggered to do (i.e., that she does habitually). 'Necessary tasks' are those tasks that the head nurse believes are strategies, that is, those tasks she believes others usually do and are expected. Being able to cover those tasks then means that the head nurse has competences that are required to do those tasks. One can thus interpret this property as 'A head nurse has the required competences to do the activities that are strategies for her and that can be triggered for the secretary.' We can query which agents have the required competences for such activities.

```
SELECT ?agent ?sharedActivity ?competence
WHERE {
  a:Secretary a:belief ?personalActivity.
  ?personalActivity a:trigger ?contextElement.
  a:Secretary a:belief ?sharedActivity.
  ?sharedActivity a:requiredCompetence ?competence.
  a:Secretary a:hasCompetence ?competence.
```

```
?agent a:belief ?sharedActivity.
?agent a:hasCompetence ?competence.}
```

The formal reasoner returns the overall head nurse, but not the team head nurse. The property is thus false in the sense that not all the head nurses can cover for the secretary. The secretary usually contacts the ER, but although the team head nurse believes this is necessary she does not have the competence to do it herself.

Property 3 We make the property more precise by specifying 'needing equipment' and 'can help each other other out'. Here 'need' means that one requires certain resources to do the actions they usually do. One way to 'help each other out' can be to usually do actions that do not use the same resource, so that it is free to use for the other. One can thus interpret this property as 'The staff usually does actions that are afforded by different resources than the actions they believe others usually do.' To specify this into a query, we query if someone actually does use a resource that is needed by others. If the query returns nothing, we know our original property is satisfied.

```
SELECT ?staffmember ?personalActivityIDo ?sharedActivityOthersDo
    ?resource
WHERE {
    ?staffmember a:belief ?personalActivityIDo.
    ?staffmember a:belief ?sharedViewOnActivityIDo
    ?staffmember a:belief ?personalActivityOthersDo
    ?staffmember a:belief ?sharedActivityOthersDo

    ?personalActivityOthersDo a:strategy ?parentActivity.
    ?sharedActivityOthersDo a:affordance ?resource.
    ?personalActivityIDo a:trigger ?something.
    ?sharedActivityIDo a:affordance ?resource.
FILTER (?staffmember != a:Patient)
FILTER (?myActivity != ?othersActivity)}
```

The first part of the query sets up two activities on which an agent has a personal and shared view. The second part specifies that one of those activities is a strategy (i.e., something others usually do) and the other one a trigger (i.e., something the staff member usually does). It then queries if a resource is used for both of these activities. The third part specifies that staff members are not patients and that the activities should be different. When the activities are the same the agents are namely cooperating using the same resource to do the same activity. The query indeed returns nothing, showing that our staff members help each other out using their knowledge of what equipment the others need.

Property 4 and 5 These properties can not be expressed in our formalism. Property 4 has a deontic flavour, that is, what one should do. Our formalism express strategies, that is, what one usually does. It does not capture that there is a consequence of not adhering to the norm. In future work, we could research if the gained

expressibility by a deontic extension of our formalism is worth the extra complexity. Property 5 mentions the length of stay, which is a common key performance indicator in operational research literature on optimizing patient flows in hospital departments. To express anything about how long the length of stay is we would need to formalize the dynamic parts of social practices. However our formalization is limited to reasoning about the static structure and so we are not able to express this property in the current terminology.

6.6. Related Work

Our work is related to other ontologies (or high-level descriptions) of agents that aim to capture sociality for social analysis. This section aims to explain that our social practice formalization is unique in its richness of expressing the social world. It expresses the social world in a unique way, namely in terms of shared action-associations: associations with competences, values, context elements or other actions. Consumat is a meta-model for consumer agents [135]. The social is captured in that an agent can see the behaviour of other agents and choose to adopt it. In other words, agents share which actions they can see, but not associations they might make with those actions. Consumat mainly focusses on what comprises the agent instead of what comprises an action. The fact that our formalization expresses the shared associations of an action is what allows us to express the social properties described in Section 6.3.3. OperA is a framework for agent organizations [73]. The social is captured in shared 'contracts'. Contracts consists of agents, roles, clauses and objectives. Agents can thus reason about the social world in these terms of these top-down prescriptions: the objectives and responsibilities of agents, instead of the bottom-up associations one makes with actions. OperA can better express deontic statements about what other should do, but cannot express the shared action hierarchy, affordances or required competence that is needed to express the aforementioned properties about the equipment others need or what others value. MAIA is a meta-model to capture agent institutions [101]. It builds on the assumption that, "while understanding and explaining individual behaviour is extremely complex, social rules or institutions are more elicitable". It places the social in these shared rules and institutions. Our social practice ontology also considers some of these rules (strategies), but relate them to one practice instead of to one model. That is, each practice has a different set of relevant strategies. MAIA's focus on institutions allows it to express deontic statements, but it can also not express the shared associations humans make with actions that allow one to express the properties about equipments others might need or what others value.

6.7. Conclusion and Future Work

This chapter aimed to formalize the social dimension of an ER to be able to automatic identify social bottlenecks. Section 6.3 presented this social dimension and Section 6.4 showed that we could capture it in a precise, unambiguous ontology based on social practices. We used the Protégé tool and a formal reasoner to ensure the coherency of the ontology. This check means that the meta-model is satisfiable

(there is a possible instance), the model is consistent (e.g., one does not claim two individuals are different and the same) and that the model is an instance of the meta-model. The OWL ontology with the meta-model and model is available online. It provides a basis for decision support tools for ER. For example, the tool could be used to identify activities where the staff does not show understanding of certain needs of the patient. Management could then educate the staff about this deficit and improve the wellbeing of both staff and patient. Our work can also be used as an example for formalizing the social dimension of other multi-agent systems. The formalization enables us to use a formal reasoner to keep track of a complex domain; something hard to do analytically. In addition, it allowed us to gain precise insights about the formal relation between different concepts (as discussed in Section 6.4).

This chapter focussed on giving a proof of concept of how to use a formal reasoner to verify a number of properties one could desire of an ER. Future work that aims to prescribe ER protocols requires a more in-depth empirical, legal and ethical study is required of the domain (see 8). Although, the ontology is limited in expressing normative and dynamic statements, we showed that the ontology is unique in its ability to verify complex social statements about helping colleagues, the needs of patients and understanding what is important. An important aspect of real social systems is that actors can deviate from established practices. Such deviations are represented in social practices, which can initially represent established practices and then evolve over time as agents enact them. The social practice model that we have shown in this chapter is static though and rather represent a snapshot of a social practice. In future work, we aim to extend the model to support evolution and expressing properties that have a time component. This chapter demonstrates that the current ontology can already be used to ensure staff and patient wellbeing by identifying possible social problems.

Acknowledgements

This work is supported by the PhD project between PDC A/S and Technical University of Denmark. We thank the institute Region H, which manages the hospitals in the Danish capital region, for providing input for the use case. We would also like to thank Helle Hvid Hansen, Jørgen Villadsen and Jordi Bieger for comments on the ideas described in this chapter and comments on a draft.

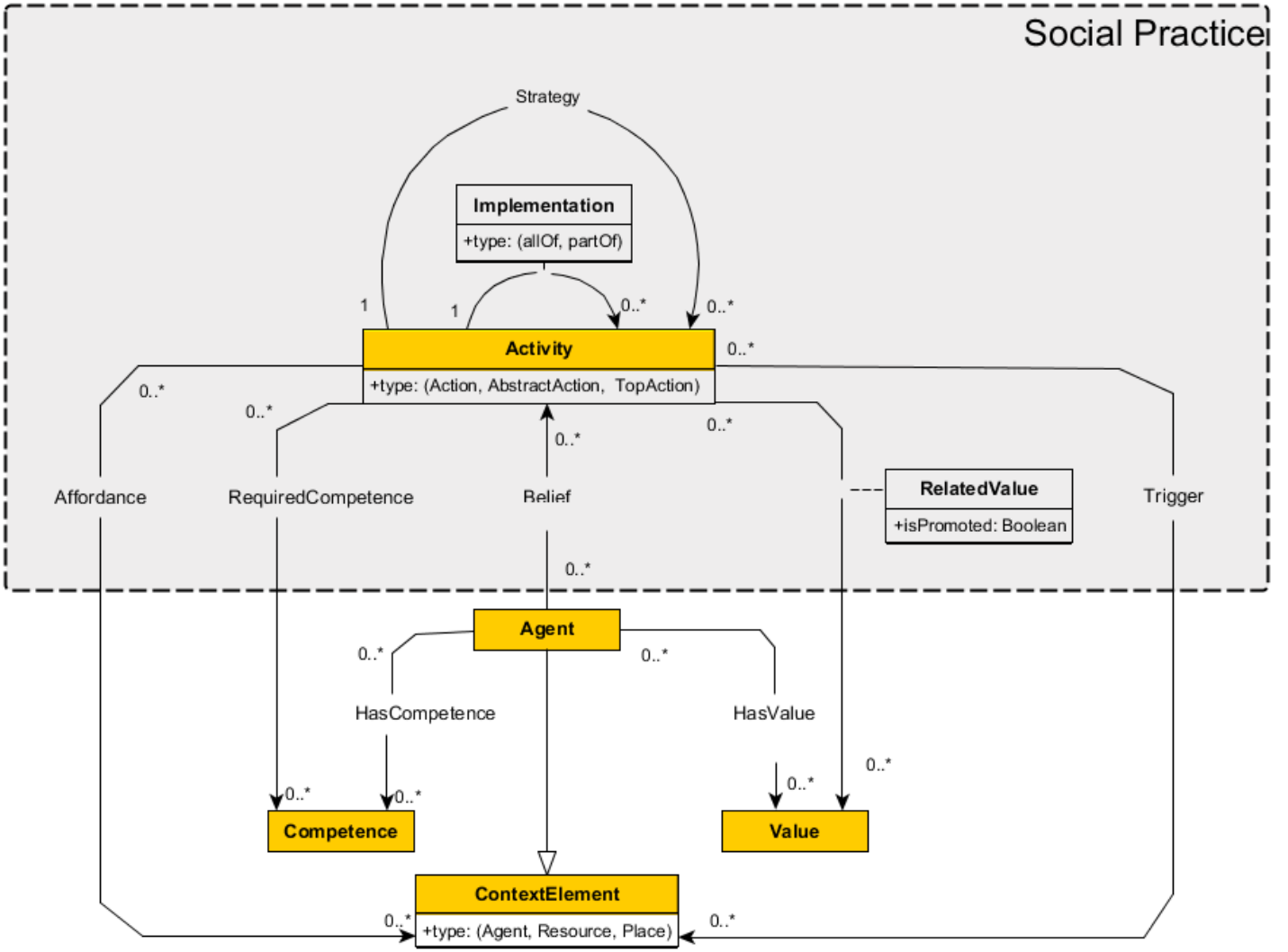


Figure 6.1: The social practice meta-model captured in the Unified Modelling Language, including classes (yellow boxes), associations (lines), association classes (transparent boxes), navigability (arrow-ends) and multiplicity (numbers).

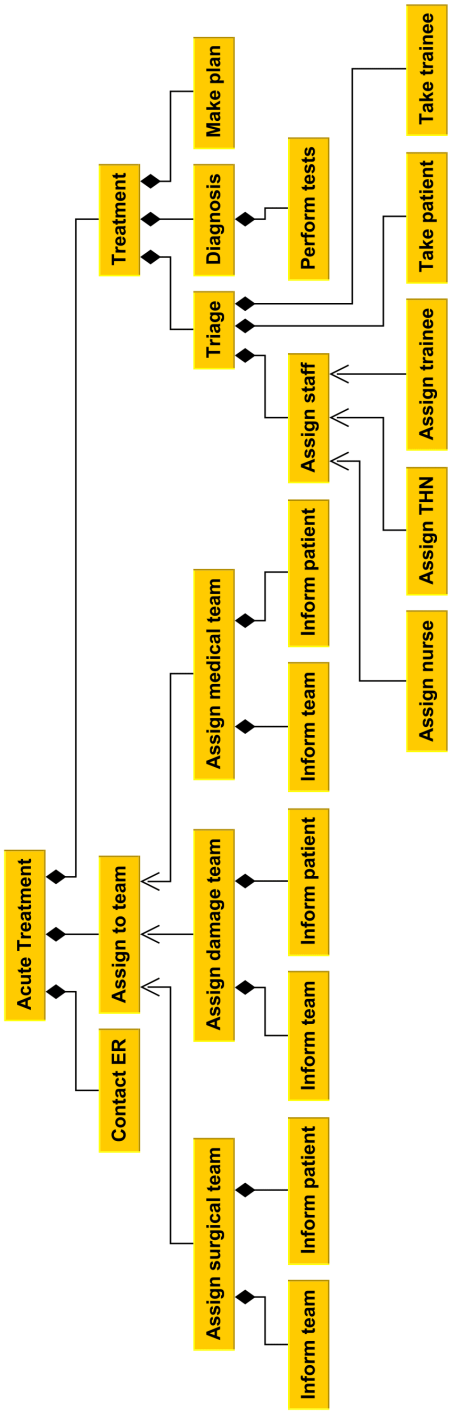


Figure 6.2: The activity tree that (roughly) represents the different phases of ER patient treatment. The arrows with a diamond end represent partOf-implementations and the arrows with an arrow end represent allOf-implementations.

7

Comparing Socio-Cognitive Theories: Discerning Two Theories on Habits

All of our life is but a mass of small habits - practical, emotional, intellectual and spiritual - that bear us irresistibly toward our destiny

William James, Talks to teachers on psychology and to students on some of life's ideals

7.1. Introduction

There is an increasing interest in considering the influence of habits on behaviour [124, 148, 154, 215]. Habitability refers to the principle that behaviour persists because it has become an automatic response to a particular, regularly encountered, context [154]. Habits have been shown to be an important driver of behaviour (e.g., in transport choices [124], food choices [215] or recycling [148]). To change behaviour it is thus important to understand how habits break [154].

Breaking habits is studied on a behaviouristic level and a cognitive level[94]. On a behaviouristic level, a habit breaks if an agent portrays different behaviour given the same context. On a cognitive level, a habit breaks if the mental connection between the context and an action is gone. On a cognitive level, a habits can thus persist even when the observable behaviour changes [267]. We will refer to a habit breaking on the behaviouristic level as 'the suspension of habitual behaviour' and to a habit breaking on the cognitive level as 'the decrease of the habitual connection'.

This chapter is accepted and in press as R. Mercuur, V. Dignum, and C. M. Jonker, "Discerning Between Two Theories on Habits with Agent-based Simulation," in Proceedings of the Social Simulation Conference 2019, 2019, p. (in press) [175]

This paper aims to compare two theories on breaking habits focusing only on their long-term dynamics. We study a scenario where an agent is first motivated to do one action (e.g., take the car) and then motivated to do another action (e.g., to take the bike). In case of a successful intervention (e.g., [3]), most agents will change their behaviour (i.e., suspend their habitual behaviour). We identify two theories in the psychological literature that can explain this dynamic on a cognitive level: the decrease theory and persist theory [3, 35, 156, 160]. The decrease theory states that a new habitual connection (i.e., the bike-habit) emerges and the original habitual connection (i.e., the car-habit) *fades out* [156, 160]. The behaviour change is a consequence of the agent enacting the new habit. The persist theory states that a new habitual connection emerges, but the original habitual connection *persists* [3, 35]. The behaviour change is a consequence of the agent intentionally choosing the new action between two (now equally strong) habits. We construct two models to compare the theories and verify these models accurately represent these theories by using simulation.

This paper shows that the two theories lead to different behaviour when the agents are motivated to do *multiple* alternative actions (e.g., take the bike or take the train), instead of *one* alternative action (e.g., take the bike). In the decrease model, the alternative action is taken up and replaces the old habit. In the persist model, the original action persists and no new habit emerges. We explain this difference using the simulation: if the original habit does not decrease, then doing multiple alternatives does not lead to the development of a strong enough habit to replace the original one. This finding is relevant for the social scientific field, because (1) it shows a scenario where it matters if habits persist (i.e., the persistence influences behaviour change) (2) it enables an empirical experiment to discern the two theories.

The remainder of the paper is structured as follows. Section 7.2 summarizes literature on habits (in particular the decrease theory and the persist theory) into properties. Section 7.3 uses these properties to construct two models: a persist model and a decrease model. Section 7.4 verifies that the models accurately describe the theories by using simulation. Section 7.5 describes the simulation experiment that shows the two theories are discernible when the agents are motivated to do multiple alternatives.

7.2. Psychological Literature on Habits

Habitual decisions are fast automatic decision that contrast with a slow intentional decisions [95, 250, 267]. Habits moderate the intention-behaviour relationship [95, 250]: the stronger the habit, the weaker the intention-behaviour relationship. For example, a strong 'car habit' weakens the influence of a 'bike intention' on behaviour. Habits predict behaviour without mediation by intentions [267]. Thus even in the absence of intention a habit continues to predict behaviour. For example, even when one does not intend to use the car anymore one can be 'stuck' in the habit of using a car. We require our habit models to separate between habits and intentions, include the moderating effect of habits on the intention-behaviour relationship and that intentions do not mediate the habit-behaviour relationship.

When a habitual decision is triggered depends on the strength of the habit and the current performance context (i.e., the context in which the agent acts) [267]. Habitual decisions are triggered by specific context-elements [267]. For example, the context-element 'home' can trigger the habit of taking the car (whereas the context-element 'work' might not). Furthermore, context-elements trigger a habitual decision only when they are nearby (i.e., part of same context as the agent). Thus, it is not so much of the habit *in general* that triggers the habitual decision, but the strength of multiple mental habitual connections specific to an activity, agent and nearby context-elements. We require our models to take into account which context-elements are part of the performance-context and with which habitual connections they are related to the deciding agent and the activity under consideration.

The amount of attention attributed to the action influences the decision to act out of habit [197]. The more attention attributed to the decision the lower the chance the action is done out of habit. The literature on the regulation of attention is extensive [10, 220]. Furthermore, to model attention one needs to take into account how different activities interact. When different activities run in parallel, conflicting or cooperating actions can influence which actions gain attention and therefore to what extent an action is done habitually [197]. Given the focus on this paper on the persist theory and decrease theory, we simplify attention and interaction with other activities to a normally distributed variable that lowers the chance the action is done out of habit.

This paper identifies two theories that can both explain the suspension of habitual behaviour, but differ in how a habitual connection updates over time: the decrease theory and the persist theory.

Decrease Theory

The decrease theory states a new habitual connection (i.e., the bike-habit) emerges and the original habitual connection *fades out* [156, 160]. The suspension of habitual behaviour is thus a consequence of the agent enacting the new habit. Lally et al. [156] showed that the automaticity individuals report decreased by an average of 0.29 (on a 7-point scale) after missing an opportunity to enact the action. This decrease is small and had no long-term effect. However, this implies habits might lose strength over time. On the individual level, the Machado's model of conditioning [160] studies how mental connection between context-elements and actions update. In the model, an association loses strength when the context-element is presented, but the action is not. The context-element activates a corresponding mental node, which in turn starts a period of 'extinction' where the association at first loses strength quickly, but then decelerates until the strength loss comes to a halt. These authors thus theorize that a habitual connection loses strength when a context-element is presented, but the activity is not enacted.

Persist Theory

The persist theory states a new habitual connection emerges, but the original habitual connection *persists* [3, 35]. The suspension of habitual behaviour is a consequence of the agent intentionally choosing the new action between two (now equally strong) habits. Bouton [35] argued for this theory when he advised that

automatic elicitation of an unwanted habitual response will likely require that the associated cue is linked with a new alternative response, rather than a non-response. Adriaanse et al. [3] showed that, at least in the short-term, performing an alternative action does not immediately replace the automatic activation of the original action with the alternative. However, once again, this leaves open the effect in the long-term. These authors thus theorize that a habitual connection does *not* lose strength when a context-element is presented, but the activity not enacted.

Both theories agree that a habitual connection gains strength when an agent performs an action in the setting of a context-element [267]. Lally et al. [155] empirically studied this strength gain in an experiment where subjects were asked to do the same action daily in the same context and report on automaticity. The subjects reported a gain in habit strength that followed an asymptotic curve and converged at a different maximum habit strength per subject. Similar results have been found when strength gain is studied on the individual level. For example, Ashby et al. [15] uses Hebbian learning to capture the strength gain of habits. Hebbian learning is based on neurology and states if two or more neurons are co-activated, the connection between these neurons strengthen [98]. In our case, this implies the habitual connection strengthens each time the action is done in presence of the context-element. The habitual connection thus gains strength when an action is done in presence of the context-element and this strength gain follows a different asymptotic curve per human.

The following properties summarize the literature on habits and will be used to construct two models to compare the theories:

1. Habits and intentions are both predictors of behaviour and interact:
   (a) habits moderate the intention-behaviour relationship
   (b) intentions do not mediate the habit-behaviour relationship
2. The decision to act out of habit is influenced by strength of a habitual connection and the current performance context.
3. Agents increase the chance to break out of a habit when they focus their attention on the decision.
4. Habits gain strength:
   (a) when an action is done in presence of a context-element
   (b) following a different asymptotic curve per agent
5. When an alternative action is performed in the same context, the original habit of the agent:
   (a) **decrease theory:** decreases
   (b) **persist theory:** does not decrease

7.3. Model

This section uses the properties from the last section to construct models that represent the persist theory and the decrease theory. Figure 7.1 presents a decision-making cycle for both the decrease model and the persist model. Both models follow a traditional agent cycle by sensing, deciding, acting and updating. The models differ only in one aspect: the decrease model weakens non-activated habitual connections while the persist model does not. The remainder of this section describes the models in more detail: the concepts necessary for both models and the different modules that form the decision-making cycle.

Concepts

To model the dynamics of habits we need to have a conceptual static model the agent uses to decide, act, learn and update. We construct a simplified version of the SoPrA model [169, 172] that focuses on habits (SoPrA-habits) in UML (Figure 7.2). We first describe the main classes in the model (i.e., `Activity`, `ContextElement`, `Resource`, `Location` and `Agent`) and then classes with a strength attribute that connect the main classes (i.e., `HabitualTrigger`, `RelatedValue`, `AdheredToValue`).

The `Activity` class models represent things an agent can do. For example, taking the bike, taking the train, taking the car or walking. The `ContextElement` class represents different entities in the environment. There are three different classes that specify (denoted with the white arrowhead) the `ContextElement` class: the `Location` class (e.g., work), the `Resource` class (e.g., a car) and the `Agent` class (e.g., a colleague). The `Value` class represents what one finds important in life. For example, environmentalism or efficiency. Lastly, the `Agent` class represents a decision-maker that chooses between the activities based on how strong it associates these activities with other classes.

The agent associates the `Activity` class with context-elements and values. First, it associates activities with context-elements by keeping track of `HabitualTriggers`. The `HabitualTrigger` represents to what extent a context-element can habitually trigger an activity. For example, it can capture that there is a strong habitual connection between being at home and taking the car. (Thus from now one we will use the SoPrA term `HabitualTrigger` instead of habitual connection to refer to the habitual connection between an action and a context-element.) Second, an agent associates activities with values by keeping track of `RelatedValue` instances. The `RelatedValue` class represents to what extent values are promoted or demoted by the activity. For example, it can capture that taking the car strongly promotes efficiency. The `AdheresToValue` class keeps track of which values the agent finds important. The agent uses these associations to habitually (based on triggers) or intentionally (based on values) choose between activities. The remaining association will be explained in the relevant decision-making modules.

Sense Performance Context

To sense the current performance context an agents retrieves a list of context-elements (e.g., locations, resources, other agents) with which it shares the `isIn-ContextOf`-association. When the model initializes the agent uses the `owns` association to determine resources it initially shares a context with and the `atHomeIn` association to determines the other agents it originally shares a context with.

Decide Based On Habits And Intentions

We separate between habitual decisions and intentional decisions (see algorithm 3). First, the agent retrieves for each activity how strongly it is habitually attached to the current context. Second, the agent compares this habit strength of these candidate activities against a threshold (the `habitThreshold` attribute in Figure 7.2). If the habit strength is lower then the threshold the agent filters the activity out. Third, based on how many activities remain the agent uses one of the following three options to make a decision. If zero candidates remain, habits have no influence and intention is used to make a decision. If one candidate remains, this decision is chosen habitually. If more than one candidates remain, intention is used to choose between these options. Note that the more attention attributed to the action the lower the chance the action is done out of habit. We model this by multiplying an attention variable with the threshold variable. Thus when attention is high (above 1) a higher habit-strength is needed to habitually trigger the action, lowering the chance an action is done out of habit. Recall, attention regulation is postponed to future work and in this model captured by the normal distribution N(1,0.25). Algorithm 3 summarizes this decision process based on habits, intention and attention. Two methods in the algorithm are explained in more detail:

calculateHabitStrength() To calculate the habit strength of each activity given the performance context and an agent we retrieve the `HabitualTrigger.strength` double for each context-element, in that performance context. We have a choice in how we combine these individual strengths into a total. For example, we can average or sum. In contrast to averaging, when summing a habit is triggered even if there are other context cue's distracting you from the triggering ones. Based on the intuition that - even in an abundance of context cue's - relevant context cue's will capture your attention and trigger habits, we choose a summation model.

intentionalDecision() Although we do not have detailed properties regarding the intentional decision, we do need an intentional model to contrast with the habitual model. To make an intentional choice we compare the activities based on a rating. To rate the activities we use the variables related to the `Value` class that SoPrA-habits provides. We calculate a `candidateRating` for each candidate activity based on how strongly an agent adheres to a value (`AdheredValue.strength`) and how strongly an agent relates an activity to the same value (`RelatedValue.strength`). The higher these two variables the higher the rating. The chance an agent chooses an action is based on this rating. For example, if the rating for walking and taking the car have a 5:1 ratio there is a 5:1 chance the agent will

walk. Instead of deterministicly choosing the highest rated candidate, we choose a chance model based on the intuition that a human intentionally varies in its actions to satisfy multiple values.

Algorithm 3: The decision influenced by habits and intentions

```
Data: Candidates - a list of activities, Agent - the agent making the
      decision, attention - random variable drawn from N(1,0.25)
List possibleCandidates;
foreach Activity AC in Candidates do
    habitStrength = calculateHabitStrength(Ac);
    if habitStrength > attention * threshold then
        possibleCandidates.add(A)
if possibleCandidates.length == 0 then
    chosenAction = intentionalDecision(Candidates)
if possibleCandidates.length == 1 then
    agent.chosenAction = candidate
if possibleCandidates.length > 1 then
    chosenAction = intentionalDecision(possibleCandidates)
```

Act

Given the focus of the paper the agent does not need to effect the environment with its actions. Acting thus retains to updating the `chosenAction` variable as described in Algorithm 3.

Strengthen Activated Habit Associations

The strength of a `HabitualTrigger` class increases when an agent performs the related action in the presence of the related context-element. We use the `habitRate` variable in the `Agent` class to decide the speed with which the strength updates. The model uses a Hebbian learning-rule to increase the strength [98]:

$$newHabitStrength = (1 - habitRate) * oldHabitStrength + habitRate * 1.$$

Weaken Non-Activated Habit Associations (Decrease Model Only)

In the decrease model, the strength of a `HabitualTrigger` class decreases when an agent performs the relevant action, but the relevant context-element is not present. We use a similar Hebbian learning-rule to decrease the strength [98]:

$$newHabitStrength = (1 - habitRate) * oldHabitStrength + habitRate * 0.$$

7.4. Verifying the Models Represent The Theories

This section verifies that the models portray the properties described in Section 7.2 and thus reflect the persist theory and decrease theory. Property 1-3 are verified

analytically (i.e., without simulation). Property 4 and 5 are verified in a simulation experiment performed on a use case model presented in Table 7.1.[1] In this experiment, the agents are initially motivated to take the car, but after tick 100 are motivated to take the train ($|alt| = 1$). We model 'motivating the agent' as doubling the rating of activities based on intentions. In addition, the amount of attention an agent focuses on the decision is temporarily increased by att_{extra} and discounted each timestep by att_{disc} until it returns to normal. We used Repast Simphony [198] to run the experiment and averaged over 50 individual runs.

Property 1a Habits moderate the intention-behaviour relationship as (1) strong habits prevent intention from influencing behaviour (see line 8-9 of Algorithm 3) and (2) weaker habits influence the intention-behaviour relation. The latter is shown by Algorithm 3 (line 10-11): weak habits will act as an initial filter on actions, but intentions still influence the final decision.

Property 1b Intentions do not mediate the habit-behaviour relationship as strong habits independently trigger action (see line 8-9 of Algorithm 3).

Property 2 As explained in the paragraph Sense Performance Context, the performance context influences the decision by triggering only relevant habitual connections (`HabitualTrigger` classes) and the strength of these habitual connections influences the decision.

Property 3 As explained in the paragraph Decide Based On Habits and Intentions, attention influences the habit threshold and consequently can increase the chance the agent break out of a habit.

Property 4 Figure 7.3 depicts the habit strength to take the car for each agent between tick 0 and 10. This shows that the habit strength of the agents follows a different asymptotic curve per agent.

Property 5 Figure 7.4 depicts the habit strength to take the car for each agent between tick 100 and 110; right after the agents are motivated to take the train instead of the car. This shows that in the persist model the strength of the habit persists and in the decrease model the strength of the habit decreases.

7.5. Finding A Scenario to Discern the Theories

By simulation a range of scenarios, we aim to find a case where the two theories show a different result. Based on a more course-grained initial exploration study we explore the following parameter settings: $|alt| \in [1, 5]$, $att_{extra} \in [2, 5]$, $att_{disc} \in [0.95, 0.99]$, $hr_{\mu} \in [0.01, 0.16]$ and $v_{\mu} \in [0.1, 0.5]$. We used Repast Simphony [198] to simulate these experiments and averaged over 50 individual runs. For each run, we calculate the difference between the number of agents that use a transport

[1] A full description of the computational model and initialization is available online (see Appendix 8.3).

Table 7.1: The classes and attributes of the use case model used in the verification and simulation experiment. The underlined attributes are parameters that are varied in the simulation experiment.

Class/Attribute	Instances	Class/Attribute	Instances
Resource	Car, Bike	Agent.habitRate	N(hr_μ,0.25hr_μ)
Location	Home, Work	Mean Habit Rate	hr_μ
Activity	takeCar, rideBike, etc.	AdheresToValue.Strength	N(v_μ,0.25v_μ)
Value	efficiency, environment	Mean of Value Adherence	v_μ
Agent	1-15	Attention Discount Rate	att_{disc}
RelatedValue.Strength	N(1,0.25)	Amount Of Alternatives	$\|alt\|$
HabitualTrigger. Strength (initiation)	0.0	Temp. Extra Attention	att_{extra}

mode in the persist model and in the decrease model (e.g., 12 agents use a car in the persist model but only 3 agents use a car in the decrease model). Next, to obtain the total difference ($\Delta_{d,p}$) between the decrease model d and the persist model p we sum over the different transport modes. This results in a performance measure $\Delta_{d,p}$ that represents the difference between the two theories for each parameter setting.

We found that for $|alt| = 2$, $att_{extra} = 4$, $att_{disc} = 0.95$, $hr_\mu = 0.01$ and $v_\mu = 0.4$ the distance between the two models ($\Delta_{d,p}$) is maximal. This represents a scenario where agents are at first motivated to take the car, but after tick 100 are motivated to do multiple alternatives: take the train or take the bike. The results are depicted in Figure 7.5. The figure shows that in the persist model agents predominantly take the car and in the decrease model the agents switch their behaviour to taking the bike or train. We explain this result by obtaining the mean habit strengths for the different transport modes from the simulation run. In the persist model, the car habit does not decline and the newly motivated behaviours (i.e., taking the train or taking the bike) do not lead to a strong enough habit to surpass the car habit. Therefore the agent habitually decides to go by car. In the decrease model, the car habit declines and the new bike or train habit surpasses the car habit. Therefore the agents adopt the new behaviour and go by car or bike. In short, in a scenario where agents are motivated to do multiple alternatives the two models show a different result: the agents adopt the new behaviour or not. We interpret this as that the decrease theory and persist theory can be discerned in an empirical experiment where humans are motivated to do *multiple* alternatives.

Using sensitivity analysis we found that the difference between the two theories in this scenario ($\Delta_{d,p}$) is depended on the mean of the habit rate (corr(hr_μ,$\Delta_{d,p}$ = -0.69) and the amount of attention given to the decision after the intervention (corr(att_{extra},$\Delta_{d,p}$) = -0.17). The mean habit rate and the amount of intervention given to the decision are variables of a different character than the number of alternatives that are motivated. The $|alt|$ is a factor that is easy to manipulate in an experiment: one treatment group is motivated to take the train whereas the other treatment group is motivated to take the train or bike. The hr_μ and att_{extra} are factors that are hard or impossible to manipulate in an experiment. Although the sensitivity to these variables cannot be used as a treatment factor, this sensitivity is

relevant to selecting the sample (e.g., selecting people that learn habits fast). The sensitivity analysis thus shows that to discern the two theories a sample is needed with subjects that learn habits fast and pay extra attention to a decision after an intervention.

7.6. Conclusion

This paper aimed to compare two theories on habits focusing only on the implications of the long-term dynamics of updating habits. We showed that the two theories lead to different behaviour when the agents are motivated to do *multiple* alternative actions (e.g., take the bike or take the train), instead of *one* alternative action (e.g., take the bike). Our finding is relevant for the social scientific field, because (1) it shows a scenario where it matters if habits persist and (2) it enables an empirical experiment to discern the two theories.

(a) The Decrease Model (b) The Persist Model

Figure 7.1: The decision-making cycle of both the decrease model as well as the persist model.

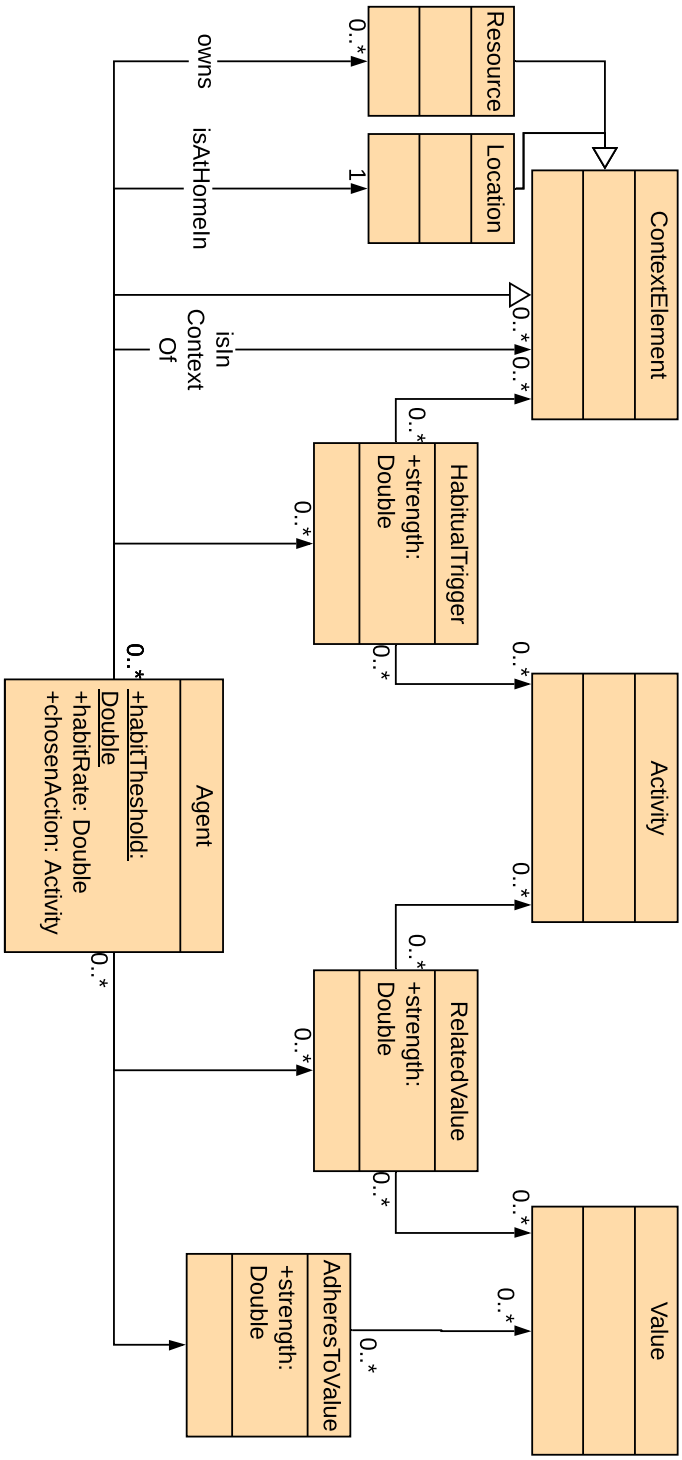


Figure 7.2: A UML Class Diagram that provides the basic concepts to model habits.

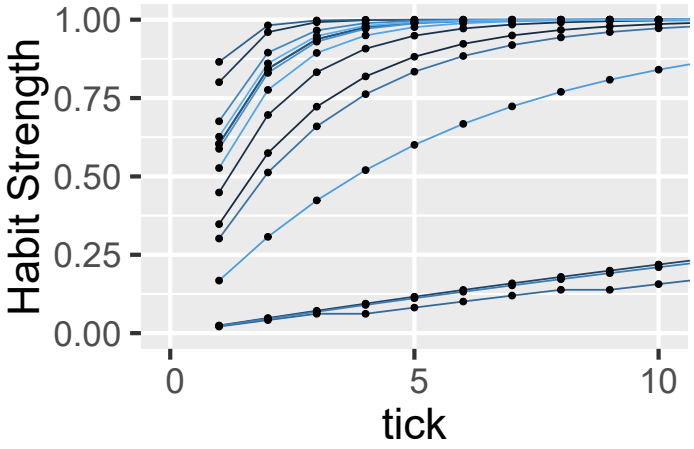


Figure 7.3: The habit-strength of 15 agents for taking the car at the onset of the simulation (tick 0-10). Each line depicts one agent.

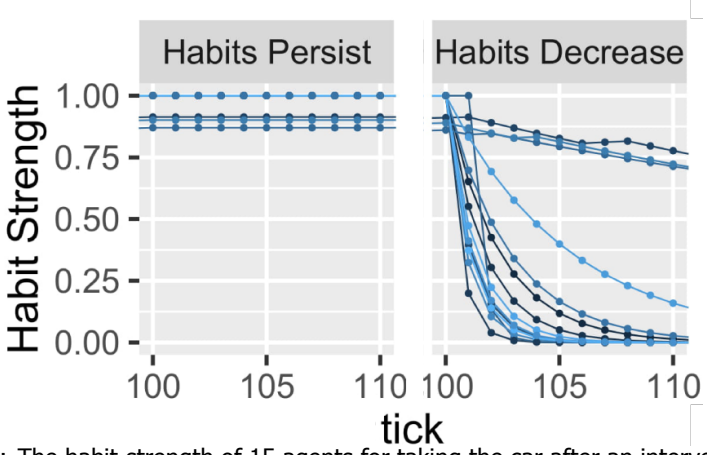


Figure 7.4: The habit-strength of 15 agents for taking the car after an intervention (tick 100-110) in the persist model and the decrease model. Each line depicts one agent.

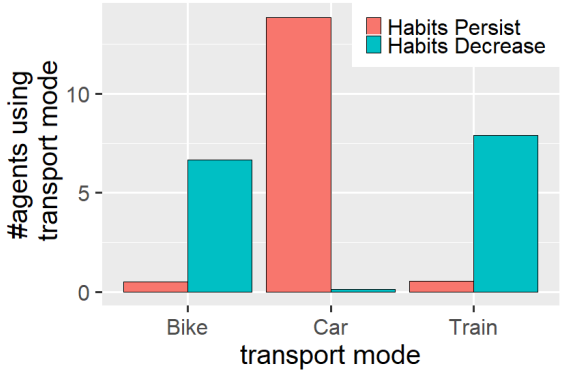


Figure 7.5: The amount of agents choosing a transport mode in the scenario where the difference between the two models is maximal.

IV

Discussion & Conclusion

8

Discussion & Conclusion

Through our efforts of self-protection, we might lose curiosity, but science IS curiosity.

If our ignorance is infinite, the only possible course of action is to muddle through as best we can.

Martin Schwartz, Why is stupidity in scientific research important?

Nothing is really solved until we understand that there is no solution. We're falling, and there's no answer to that. We can't control it.[...] The fall is a great blessing.

Joko Beck, Nothing Special

This chapter discusses the central choices made during the project, as well as the scientific and societal relevance of this thesis. We conclude with a summary of this thesis in relation to our research questions.

8.1. Discussion on Criteria, Design Choices, Framework and Use Cases

8.1.1. Criteria

In the creation of SoPrA, we emphasised two criteria that align with the motivation of ABSS: SoPrA's reusability and SoPrA's grounding in evidence (see Chapter 1). SoPrA promotes reusability by combining domain-independent concepts in a modular and parsimonious framework that supports multiple domain-specific models (see Chapter 4 and Part III). Recall that we differentiate between a framework (domain-independent) and a model (specific for a study and domain). SoPrA promotes grounding by adhering to requirements based on evidence on routines from social

psychology, sociology and agent technology (see Chapter 3 and 4). By emphasising grounding and reusability, SoPrA supports ABSS researchers in understanding social systems through a collective scientific endeavour (see Chapter 1).

In retrospect, we gained more clarity on how the wording 'grounding in evidence' over 'realism' aligns with our methodological view. Realism implies the positivistic view of aiming for the one right model that reflects the one right reality [54]. Our view is interpretivistic: there are many different perspectives on reality that all reflect an interpretation of a social system. These interpretations are not absolute (i.e., we did not make the only possible model) nor arbitrary (i.e., not any interpretation is true). Instead, SoPrA is one perspective on human behaviour grounded in one class of socio-cognitive evidence, namely evidence on routines from social psychology, sociology and agent theory. For example, our conceptualisation of habitual connections is not absolute (e.g., neurological models also show a useful perspective) nor arbitrary (e.g., our conceptualisation integrates evidence that the gain in habit strength follows an asymptotic curve [155] – not just any random curve). Thus, our criteria align with our view that multiple views that are founded in evidence complement each other.

By working with grounding and reusability, we disovered that these criteria promote other criteria. First, by grounding SoPrA in evidence, we met criteria of empirical grounding (e.g., the strength of a habit is dependent on context) as well as theoretical grounding (e.g., we enable context-dependence by modelling different context-action relations) (see 4.3). Second, by ensuring SoPrA is (re)useable, we focused on a clear and parsimonious framework. Parsimony reduced the number of variables making the model easy to interpret (see Dignum et al. [72] on the necessity of parsimony when critiquing ACT-R and 4.6 for a discussion on the parsimony of SoPrA). Last, we found other criteria are met because we combined grounding and reusability in one framework. For example, by increasing the reusability, and consequently, the parsimony of a framework grounded in evidence, we improved the generality of SoPrA. That is, if we only have a few mechanisms in SoPrA that still explain a wide array of evidence, this shows the mechanisms are fundamental and generally applicable. Thus by emphasising both reusability and grounding SoPrA became more general. In short, by focusing on grounding and reusability, we met other criteria that were subsumed by either grounding (e.g., empirical grounding) or reusability (e.g., parsimony) or by the combination of grounding and reusability (e.g., generality). We advise future research to further investigate the relationship between the different criteria used to evaluate agent-based models.

Creating SoPrA required balancing reusability and grounding. By creating a (re)useable framework, we opened up new engineering possibilities. By creating a grounded framework, we limited the number of possible models to those that are in line with the evidence.[1] An example of a design choice that requires balancing is related to requirement 5 for our dynamic extension on habits in Chapter 7: Habits

[1]This balance between reusability and grounding represent two sides of constraints: they empower and limit [202]. In mathematics, in particular logic, a formal language aims to balance the number of constraints on the language in order to be both expressive and ensure certain properties [128, 258]. In design sciences, a creative process is stimulated by balancing the number of constraints– neither limiting the creative process too much nor opening it up so much one gets lost in space [27, 202].

gain strength following a different asymptotic curve per agent. From the reusability perspective, we want to relax this requirement to allow a wide range of curves. On the other hand, from the grounding perspective, we want to sharpen this requirement to match the evidence that the maximum of this curve is reached on average in 66 days [155]. The current requirement balances these design criteria by restricting both the number of models and allowing flexibility. SoPrA aims for a balance between grounding — in the sense of restricting possibilities to match evidence — and reusability — in the sense of opening up new engineering possibilities. The next section discusses how this balance between reusability and grounding is reflected in our choices for certain behavioural aspects and their implementation.

8.1.2. Design Choices: Behavioral Aspects, Requirements and Framework

The creation of SoPrA required selecting certain behavioural aspects (RQ2), requirements that reflect current evidence on those aspects (RQ3) and designing a framework that adheres to those requirements (RQ4). These design choices were made based on our evaluation criteria (i.e., grounding and reusability), our aim (i.e., to model agents with routines) and our audience (i.e., ABSS researchers).

HSI-aspects

We found that the habitual, social and interconnected aspects (HSI-aspects) of behaviour are important for ABSS and are emphasized by SPT (see Chapter 3 and Chapter 4). First, our literature review shows the HSI-aspects are emphasized in SPT (see Chapter 3). Second, the review shows the HSI-aspects are relevant for ABSS: they reflect current evidence that people's decisions are automatic and contra-intentional (see Section 3.2.1), based on human values (see Section 3.2.1), based on a collective view on the world (see Section 3.2.2) and depend on hierarchies of activities (see Section 3.2.3). In addition to being grounded and relevant for ABSS, the HSI-aspects resulted in a modular framework that served our aim to model agents with routines (see Chapter 4).

We advise future research to further investigate the relationships between routines, the HSI-aspects and other aspects described by SPT. For example, other aspects that SPT emphasises are usual behaviour, shared behaviour, sequenced behaviour, everyday behaviour, custom behaviour, normal behaviour, best practice behaviour, know-how behaviour and role behaviour (see Chapter 3). Given our literature review (Chapter 3), comparison with other possible frameworks (3 and 4), and applications III, we postulate that the HSI-aspects suffice to describe routines. Chapter 4 describes how the HSI-aspects suffice to describe several other aspects of routine behaviour. For example, we discuss how SoPrA expresses a plan pattern (a way to model sequenced behaviour) as an interconnection of habitual actions that are socially shared (see Section 4.5). Proving that the HSI-aspects suffice to describe *all* aspects of routines is to proof the inexistence of the proverbial 'black swan'.[2] That is, we can only show that, to this date, given our review, model com-

[2]Another way to argue for the sufficiency of the HSI-aspects is to look into the etymology of routines. The word 'routine' comes from the metaphor of a beaten path (a 'route'); a path that is beaten over

parisons and applications, we found no counter-example. Future research will gain additional clarity on the sufficiency of the HSI-aspects by applying SoPrA to simulate routines and evaluating if SoPrA suffices. SoPrA enables such investigation by providing a computational model of the HSI-aspects to support these simulations.

We advise future research to investigate extensions of SoPrA with aspects of behaviour that fall outside to scope of routine behaviour. Chapter 3 describes how there are aspects emphasised by SPT that fall outside routine behaviour. For example, competencies are a precursor to routines and not necessary to describe routines themselves. In addition, Chapter 3 describes HSI-behaviour that falls outside the scope of routine behaviour. For example, non-routine aspects of social behaviour comprise emotional behaviour [111], social networks (including trust and reputation) [140, 195] and chains of beliefs as modelled by the theory of mind [211]. Appendix 8.3 shows a possible extension of SoPrA that includes competences, affordances and roles. This UML model serves as the first step towards a systemic extension of SoPrA that includes a literature study, requirement elicitation and transparent model choices. In short, SoPrA focuses on those aspects of SPT that suffice to model routines and provides an anchor in the future investigation of aspects that fall outside the scope of routines.

Requirements

Requirements elicitation resulted in a clear relationship between SoPrA and current evidence (see Chapter 3 and Chapter 4). First, section 3.2 presented our requirements and their relation to current evidence on routines from agent theory, sociology and social psychology and as such clarifies which evidence is taken into account. Second, our requirements provided a structure to evaluate current agent frameworks and to show these frameworks do not suffice to capture routines (Section 3.3). Last, Chapter 4 uses our requirements to evaluate SoPrA and show it represents current evidence on routines.

We advise future work to add, change, or remove requirements based on new evidence or new extensions of SoPrA. We hope the transparency that requirements bring helps researchers in a systematic improvement of SoPrA. Future research that explores a dynamic extension will need requirements that describe the evidence on the dynamics of routines. Chapter 7 shows examples of such requirements for habitual behaviour. As shown in Chapter 4, the requirements for the static framework we provide are captured by design. For complex requirements that describe dynamic properties of routines, we advise the use of formal methods or ABSS to aid the verification.

Model Choices

Chapter 4 describes the model choices necessary to ensure SoPrA adheres to our requirements and, in a broader sense, to serve our aim for a grounded and reuseable framework to simulate routines. In particular, we argue for the choice to enable a connection between each activity, element and agent (Section 4.2), the choice to use habitual triggers as a primary concept in the model (Section 4.3), the choice

time (habits), by different people (social) and through different activities (interconnectivity).

to enable multiple collective views on an SP (Section 4.4.1) and the choice to use the same concepts to model the agent's personal view and collective view (Section 4.4.2).

This thesis describes a process of organising and labelling concepts and relations until we found our grounded and reusable framework. The progressive modelling insights are reflected in the differences between earlier versions of SoPrA (as found in Chapter 8.3 and 6) and the final version in Chapter 4. For example, in Chapter 8.3 norms are modelled as a stand-a-lone relation with no clear relation to habits. In contrast, Chapter 4 models a norm as a social version of a habitual trigger. As argued, this modelling choice is necessary to represent the evidence on the semantic connection between norms and habits and improve the modularity of SoPrA. We hope that the transparency and systematicity in our design choices enables future research to build upon our work. We advise future work to use SoPrA to analyze the co-linearity between the proposed variables and the concepts.

8.1.3. Conceptual, Formal and Computational Aspects

The conceptual aspects of SoPrA are described using UML, SoPrA's formal framework is defined in the OWL-based Protégé, and a computational framework has been implemented in Repast Java (see Chapter 4).

UML Figure 4.3 provides a conceptual framework in the UML-language. UML is the (best practice) standardised modelling language that makes SoPrA easy to read, understand, code, extend, maintain and adapt [25]. This serves our aim to make a (re)useable framework and suits the ABSS community; a community that consists of both scientists with a background in computational sciences and scientists with a background in social sciences.

OWL Our online repository provides a formal framework in the OWL-language. Chapter 8.3 and Chapter 6 discuss the translation from the UML model to the OWL model and discuss its advantages. In short, a formal framework promotes unambiguity, rigour and enables formal reasoning. Using Protege – a formal reasoner – we show that SoPrA is consistent, verify that our domain models correctly implement SoPrA and verify properties (see Chapter 6).

Java Our online repository provides a computational model in Repast Java that enables simulating social phenomena with SoPrA. Java is a general-purpose computer language and profits from a large database of packages to use in combination with SoPrA.

In short, the conceptual, formal and computational framework all have different advantages and complement each other.

Current research in the agent literature focuses on providing a formal semantics for existing agent frameworks (e.g., 2APL, GOAL). A formal semantics offers a complete and systematic translation from the syntactical language to a language that describes the semantic models. For example, by translating the syntactical language to Kripke models, truth tables, or to another language for which the formal semantics has already been defined [112]. Formalizing SoPrA promotes the

unambiguity of SoPrA and enables formal reasoning. Our OWL-framework has a formal semantics that focuses on ontological relations (e.g., subsumptions, equivalence). By using other formal languages (e.g., epistemic or dynamic), future research will be able to promote unambiguity and enable formal reasoning on these (e.g., the epistemic or the dynamic) properties of routines. Note that to provide a formal semantics that correctly represents routines, it's first necessary to define the underlying concepts clearly (e.g., needs, values), and discuss their relationship. Chapter 4 describes such a discussion and provides the basis for future work on a dynamic or epistemic formal semantics of SoPrA.

8.1.4. Reflection on Use Cases

SoPrA (in some form or another) has been applied to the ultimatum game (Chapter 2, Chapter 5), commuting (Chapter 4, Chapter 7, Roes [217]'s master thesis), rumourmongering (Chapter 8.3), the emergency room of a hospital (Chapter 6), PV-installation (D'Haens [65]'s master thesis), value-alignment of AI (Chapter 5) and meat-eating (my master thesis [167]). This wide array of domains gave us insight into the design choices mentioned above and the scope and relevance of SoPrA (see Section 8.2). A first insight relates to the danger of using SoPrA as a golden hammer (i.e., "If all you have is a hammer, everything looks like a nail"). We found that due to the prevalence of routines in many domains the applicability of SoPrA is wide. However, the PV-installation and ultimatum game case studies showed that when behaviour is one-off or heavenly restricted SoPrA is ill suited to study such domains. A second insight builds upon the first and relates to the ultimatum game (UG). In the UG, decision-making is studied within the artificial setting of a strict lab game. As discussed in Section 2.6, lab experiments have the advantage of allowing measurements in a reproducible controlled setting (which is the reason we were able to obtain a relatively large homogeneous dataset of a meta-study). The lab experiments have the disadvantage that this setting is unrepresentative for real-life decision making [147]. In particular, in the UG where behaviour is both new (due to the game being new) and restricted (only certain actions are allowed) agents cannot fall back on old habits. This restriction meant we had to find different domains (e.g., commuting) to model all aspects of routines. In hindsight, the restriction of the UG is also an advantage: they enable to test parts of the model in an artificial setting where habitual behaviour plays a minor role. As such, we found that agents often repeat behaviour even without habitual behaviour in place. We encourage future work to use such restricted settings to their advantage and build upon our results to further compare the relative role of habitual, social and interconnected behaviour in routine behaviour.

8.2. Relevance

Working responsibly as a scientist requires that we consider not only the scientific relevance of scientific work but also its potential impact on society.

8.2.1. Scientific Relevance

The scientific relevance of SoPrA is that it makes routine behaviour explicit and computational. In combination with ABSS, this supports researchers in understanding the role of routines in social systems. As we have shown, routine behaviour can be modelled explicitly in terms of habitual, social and interconnective behaviour. The remainder of this section describes the relevance of our work for modelling those types of behaviour.

Habituality

SoPrA enables modellers to endow agents with habitual behaviour and intentional behaviour based on values. SoPrA fits in a wider transition from a goal-oriented view on humans to a more balanced view of human behaviour as both habitual and intentional. As argued in the introduction, this transition is necessary as not all human behaviour aligns with our goals and values [124, 136, 215]. SoPrA extends current frameworks by supporting context-dependent habits, individual learning concerning habits and explicit reasoning about habits (see Section 3.3). Chapter 7 illustrates how SoPrA provides insights into habitual behaviour via a simulation study that discerns two theories on habits. In short, SoPrA enables ABSS studies that help understand the role of habits in social systems and improve those systems.

SoPrA enables agents with intentional behaviour on values. Agents focussing on values instead of their welfare reflects a transition from the myopic view of the homo economicus (as found in classical game theory [190], classical economics [139] or choice models [164]) to a more diverse view on human decision-making. SoPrA enables an understanding of social systems that goes beyond a view of humans as merely self-interested and acknowledges humans interest in safety, power, tradition, and, other humans (see Chapter 2 and 4.3.4). In addition, values enable modellers to relate different mechanisms, both within the agent and between domains (see Part III). Chapter 2 and Chapter 5 illustrate how SoPrAs enables new explanations of human behaviour based on values. In short, SoPrA enables ABSS studies were agent behaviour is based on diverse values and traceable to universal, trans-contextual, moral and operationalised grounds.

Sociality

SoPrA enables agents with social behaviour. SoPrA fits in a wider transition from agents with norms to agents with SPs. Sociality, as captured in our conceptualisation, goes beyond a conscious decision to conform to norms or values [214]. Instead, sociality happens from the inside out: it is a fundamental part of how we view the world (see Chapter 3.2.2). SoPrA extends current agent frameworks and uses a comprehensive set of collective concepts, orders information around actions and relates individual and collective concepts in order to guide interaction (see Chapter 3). Roes [217] illustrates the scientific relevance of SoPrA for social behaviour. In her master thesis, she uses a simplified version of SoPrA to evaluate the difference between standard social imitation models (i.e., copying the actions of each other) and social influence as captured by SoPrA (i.e., copying each other's beliefs about actions). She finds that when the size of the group of agents increases the two models of sociality diverge. This finding implies that incorporating evidence

on sociality enables different explanations of social systems. In short, SoPrA enables ABSS studies that help understand the role of sociality in social systems and improve those systems.

Interconnectivity

SoPrA enables agents with interconnected behaviour. Interconnectivity is a necessary ingredient of decision-making as plans or protocols dictate our day. SPs show how decisions need to fit in our own standard schedule and a socially accepted schedule (see Section 3.2.3, Section 3.3 and [37, 130, 208, 225]). SoPrA extends current agent framework as it supports hierarchies of activities, explicit relations between activities and other concepts of the model (e.g., values and context-elements) and capturing how these relations are socially shared (see Chapter 3). These concepts and relations provide a basis for agents to reason about the resources they use and coordinate their plans (see Chapter 3 and 4). Although not simulated within the scope of this thesis, the influence of inter-connective behaviour on decision-making is extensive (see Chapter 3). We advise future work to simulate this key aspects of behaviour using the SoPrA framework.

Relevance for Social Science and Multi-Agent Systems

We hope our work has relevance outside ABSS for social science and multi-agent systems (MAS). As discussed in Chapter 1 and 3, both fields call to integrate habituality [104, 144, 234, 240] and sociality [43, 78, 131, 145, 196, 204, 226, 235] in decision-making. Our work is relevant for the social sciences as it (1) enables research on interventions based on an evidence-driven view on human routines (2) enables the study of human behaviour using simulation (as demonstrated in Chapter 7). Our work is relevant for engineering agents (MAS) as (1) the efficiency of human routines could improve the efficiency of agent decision-making and (2) SoPrA provides autonomous agents with a mental model to understand and interact with human routines.

8.2.2. Societal Relevance

This thesis contributes to the theoretical work necessary to support ABSS researchers in understanding the role of routines in social systems. The shortest route from SoPrA to societal relevance goes via ABSS-based policy advice; other indirect routes go via supporting models that illustrate, describe, explain or theorise social problems (see Box 1.1.1). This subsection first describes the societal relevance of SoPrA based on our simulation studies on transport-mode routines, animal consumption routines, social bottlenecks in emergency rooms (ERs) and value-alignment in AI (see Part III).[3] Second, we reflect on the general steps necessary to use SoPrA-based simulation studies for ethical policy advice and encourage future work to explore several advantages SoPrA offers in policy modelling.

[3]The sections on animal consumption routines and transport mode choice incorporate two master thesis based on SoPrA [167, 217].

Transport Mode Routines

Transport-mode decisions (e.g., to take the bike, train or car to work) affect our health, climate and economy [124]. Empirical and theoretical evidence shows the importance of routines for transport-mode choices (see Chapter 3 and [217]). For example, Hoffmann et al. [124] shows that transport mode habits have a moderate-to-strong correlation with transport mode behaviour. SoPrA focuses on integrating this evidence on routines and contrasts with standard models used in traffic mode choice (e.g., canonical choice models [163, 164] that capture traffic mode choices as an intentional decision made on a list of criteria (e.g., distance and cost)). In short, by integrating evidence on routines, SoPrA enables modellers to take into account the influence of routine behaviour in our transport mode choices.

SoPrA has been applied by Roes [217] who used an ABSS to understand the transport-mode behaviour of a group of students. She found that the intersubject sharing of routines makes it so that large groups have a higher chance to cluster towards the same behaviour than smaller groups. These findings imply that to change the behaviour of large groups, policies need to target the group as a whole (e.g., by influencing leading figures in the group) instead of targeting every individual equally (e.g., by providing individual cost benefits to reduce the cost for biking). This insight can help policy makers to change travelling routines (e.g., to mitigate climate change) and encourage future work to use SoPrA to gain similar insights.

Animal Consumption Routines

Our consumption routines drive both factory farming and climate change [47]. If the world went vegan, this would save 2 trillion fish per year, save 69.4 billion land animals per year [85], save 8 million human lives by 2050, reduce greenhouse gas emissions by two thirds and lead to healthcare-related savings and avoided climate damages of $1.5 trillion [109]. SoPrA enables the simulation on aspects of consumption routines that current agent frameworks do not capture (see Chapter 3). For example, the standard framework for consumption behaviour – the Consumat [135] framework – does not capture social aspects of animal consumption. For instance, the Consumat does not capture a collective view on animal consumption as healthy that mitigates behavioural change. Nor does the Consumat capture interconnectivity or context-dependent habits (see Section 3.3). We envision simulating these aspects will provide new insights. For example, in Mercuur [167], we used a simulation study to show that organising a 'vegetarian week' motivates a participating agent to break out of its animal consumption habit. SoPrA enables an extension of this work that includes the social and interconnected aspects of animal consumption routines (in addition to the habitual aspect). This extension enables research on the connection of animal consumption routines with shopping and cooking routines and the interplay between animal consumption habits and the collective view on animal consumption. Using an ABSS based on SoPrA we can research whether these interactions explain current animal consumption behaviour. For example, to what extent do current cooking routines prevent meat-eaters from eating plant-based? Furthermore, to what extent does a collective negative view on plant-based consumptions (e.g., unhealthy) influence individual beliefs on animal

consumption? We envision that such insights on animal consumption routines help construct effective governmental interventions that mitigate animal consumption (see Section 8.2.2).

Social Bottlenecks in Emerency Room Routines

Understanding the impact of activities on ER patients and ER staff would improve the well-being for both of them [162]. Chapter 6 uses SoPrA to formalise ER routines to identify social bottlenecks. Using a formal reasoner on a (synthetic) database based on the Herlev Hospital in Denmark, we check whether certain social requirements are met. For example, whether the head nurse can cover the tasks of a secretary in her absence. The social bottlenecks identified by the formal reasoner inform interventions that improve staff cooperation and, subsequently, patient well being. For example, the formal reasoner identified that trainees view triage as an educational activity, although this might cause undesired results in times of crisis. We find SoPrA enables a stronger focus on the staff and patients' interaction than previous models that focus on patient flow [219]. SoPrA thus provides a tool to identify social bottlenecks in organisational routines to improve patient and staff welfare.

Chapter 6 provides a tool that should be supplemented with an in-depth empirical study to identify said social bottlenecks in real-life scenarios. The model presented in Chapter 6 describes the emergency room based on a small set of qualitative observations made in the Herlev Hospital Denmark. The values, protocols, agents, competencies are described based on these (limited) observations and serve as a proof-of-concept and not as in-depth empirical evidence. If the model is used to prescribe new emergency protocols, an in-depth study is required on (1) the accurateness of the empirical observations used in this proof-of-concept, (2) the legal consequences of the proposed protocols and, (3) the values that are at conflict in an ER and their prioritisation (i.e., ethics). Bruntse [39] targets the first of these three aims by extending our model to a full-fledged ABSS informed by event log data. We encourage future work to extend Chapter 6 and Bruntse [39] to give legal sensitive and value-driven advice to improve staff and patient cooperation in the ER.

To use SoPrA for new insights in organisational routines (such as those of the ER), we advise researchers to draw from both SoPrA and organisational agent frameworks. Organisations and routines overlap in facilitating interactions between people with a shared purpose [73]. The OperA framework captures this shared purpose in agent organisations in shared contracts [73]. Shared contracts represent top-down representation of agents, their roles, objectives and their clauses. In contrast to prescribing agent behaviour, SoPrA describes the bottom-up emergence of collective beliefs based on individual beliefs. Furthermore, SoPrA offers capturing aspects of routines not captured in organisational frameworks: activity hierarchies, related habits and values. A synthesis of both agent frameworks requires extending SoPrA with a `Role` class and deontological norms. A first step towards the first is described in Appendix 8.3 a step towards the latter should be based on [69]. The integration of these concepts requires a systemic extension of SoPrA that includes

a literature study, requirement elicitation and transparent model choices. As mentioned in Section 8.1.2, SoPrA provides a modular and orthogonal anchor for such systematic extensions.

Value-Alignment in AI

As AI becomes more pervasive in our daily lives, there is a growing need to reduce AI risks for our society [9, 116]. For this, we need to ensure that the decisions and actions made by AI-systems align with human values [224]. Chapter 5 aims to train autonomous agents (AAs) such that they are equipped to estimate human values from human behaviour. The SoPrA framework is used to provide a training environment for the AA. This training environment provides control and access to the relation between the (simulated) human behaviour and the underlying (simulated) values. We found two methods that provide an increase in our confidence in the AAs estimation of human values. The methods differ in their applicability, the latter being more efficient when the number of interactions with the agent has to be restricted. This chapter thus provides insights relevant to the successful estimation of human values by AAs. More generally, we hope SoPrA contributes to value-alignment in AI by providing a simulated training partner that reproduces routine human behaviour.

Ethical Policy Advice

ABSS supports the formulation and evaluation of policies [108].[4] Using models to influence policies requires using the model in a complete policy chain: agenda setting, policy formulation, implementation, evaluation and adoption [21]. Future work should be directed towards such integration of SoPrA in practice. This dissertation focuses on the theoretical groundwork necessary to support ABSS-based policy advice. The relevance of such theoretical groundwork lies in SoPrA's ability to support current best-practices in ABSS-based policy advice: (1) making the implicit explicit, (2) providing understanding over numbers, (3) involving stakeholders, (4) communicating complex systems and (5) reflecting on the ethical implications of the model [108].

We hope SoPrA provides several advantages in such an approach:

1. SoPrA provides explanations based on domain-independent concepts (e.g., habits, collective view, context-elements), making it easier to compare results from several policy studies.

2. SoPrA is grounded in evidence from the social sciences, which improves the trust stakeholders should put in SoPrA-based models.

3. The parsimonious UML presentation of SoPrA supports the communication and understandability of policy models.

[4] Recall that an ABM provides a common language for policy modellers to formulate their policies; and simulations enable policy modellers to evaluate implications of their policies cost-effectively and flexibly [108].

4. SoPrA focuses on socio-cognitive aspects relevant for considering (the ethics of) involving multiple stakeholders (discussed below).

These advantages of SoPrA should be utilised in future work that aims to use SoPrA for policy advice.

Policy modellers that use SoPrA should reflect on ethical considerations regarding their policy advice. SoPrA equips modellers to consider three key ethical questions when modelling policies [142]:

what is (un)desirable behaviour? A principal consideration in the 'what'-question is the gap between what the individual finds desirable and what the group finds desirable [142]. SoPrA enables to explore this dynamic by making explicit (1) what agents themselves value, (2) what agents believe the group values, (3) what the group actually values and (4) what the relevant ethical values are.

why do people persist with undesirable behaviour Concerning the 'why'-question, ethical behavioural change requires to see undesirable behaviour (1) as explainable and (2) contextual [142]. As described in this dissertation, behaviour is not simply bad but aims for certain values (that might clash with societal values) and depends on a habitual and social context. For example, using SoPrA one is able to explain car-driving as a product of the availability of cars over trains (i.e., captured in the `ContextElement` class) and a product of what the group believes to be important (captured in the `myCollectiveView` attribute in the `ValueConnection` class and `ValuePriority` class) (see Chapter 4).

how to narrow the gap in an ethically acceptable way? Concerning the 'how'-question, ethical behavioural change requires the preservation of individual autonomy and involvement of stakeholders in the intervention [142, 260]. SoPrA helps capture individual autonomy by making the balance between the individual and collective perspective explicit. Furthermore, the value-based approach in SoPrA is useful for stakeholder participation as it allows stakeholders to find common values [207] and, from there, widely supported solutions.

In short, we advise future work to utilise SoPrA's ability to capture the habitual, social and value-sensitive context of policy advice.

8.3. Conclusion

This thesis provides the SoPrA framework that enables the simulation of human routines (see **RO**). The first part of the thesis identified the aspects of social practice theory that are relevant for agent-based simulation, distilled requirements from the literature, reviewed current agent models and provides the SoPrA framework that satisfies said requirements.

Chapter 2 provides a preliminary simulation study that focussed on two aspects of routines: values and norms. We found that human values provide a universal, moral, operationalised, empirically grounded and a trans-contextual basis for intentional behaviour norms provide the dynamic social component for behaviour. The chapter provides the necessary submodels for SoPrA and gives insight into what SPs capture over and above values and norms.

Chapter 3 zooms out again and considered the SP as a whole. First, we showed that the habitual, social and interconnected aspects of behaviour are both emphasised by SPT and important for ABSSs (see **RQ1**). These aspects reflect current evidence that people's decisions are automatic and contra-intentional, based on human values, based on a collective view on the world and depend on hierarchies of activities. Second, we presented our requirements for an agent framework that integrates SPT to enable simulations of human routines (see **RQ2**). Requirement elicitation resulted in a clear relationship between SoPrA and current empirical and theoretical evidence and ensures SoPrA provides a transparent anchor in further investigation. Third, we showed that current agent frameworks do not enable the simulation of human routines (see **RQ3**). In particular, current agent framework lack (1) support for context-dependent habits, individual learning concerning habits and explicit reasoning about habits; (2) a comprehensive set of collective concepts, an efficient ordering of social information around actions and a connection between individual and collective concepts; and (3) explicit relations between activities, between each activity and each other model concept and hierarchies of activities.

Chapter 4 provides the domain-independent SoPrA framework that satisfies our requirements to simulate human routines (see **RQ4**). For each modelling choice, we presented an overview of the relevant concepts and argued for a specific conceptualisation that leads to a reuseable grounded framework suitable for ABSS. In particular, we motivated a conceptualisation that enables a connection between each activity, element and agent, uses habitual triggers as a primary concept in the model, enables multiple collective views on an SP and uses the same concepts to model the agent's personal view and collective view. The model is presented in UML to enhance modularity and accessibility, implemented in Repast to enable ABSS and formalised in Protege to ensure its consistency.

This results in a framework that integrates socio-cognitive theory and agents, correctly depicts these theories, implements new and relevant behavioural aspects, computational implementable and supports simulation studies. Or in short, a domain-independent agent framework that is (re)useable and grounded in evidence.

The second part of this thesis described applications of SoPrA (**RQ5**) and demonstrates the scientific and societal relevance of this thesis.

scientific relevance This thesis provides a grounded and (re)useable agent framework that supports the simulation of routines and as such is relevant for current work in

SPT in ABSS by emphasising domain-independence, integrating agent theory and SPT and conceptualising SPT specifically to support ABS.

ABSS by enabling simulations where humans get stuck in behaviour that does not align with their goals (i.e., habits), make decisions based on more than their own welfare (i.e., values), reason based on a collective view on the world (i.e., sociality) and make decisions within a wider context of interconnected decisions (i.e., interconnectivity) (see Part III and master theses D'Haens [65], Mercuur [167], Roes [217]). By combining the strengths of ABSS and socio-cognitive theory, we enable a new way to know, explore and improve the world grounded in evidence.

MAS, AI by enabling agents with traceable human-like decision-making. This leads to autonomous agents that use human routines for efficiency and enables them to understand and interact with human routines (see Chapter 5).

Sociology and Psychology by crystallising socio-cognitive theories and enabling exploration of these theories via simulation (see Chapter 7).

societal relevance This thesis contributes to the theoretical work necessary to support ABSS researchers in understanding the role of routines in social systems. Concretely, SoPrA is relevant for:

AI safety by providing a simulated training partner for autonomous agents that reproduces routine human behaviour and helps them to successfully estimate human values (Chapter 5).

Social Routines in the ER by providing a formal model to identify social bottlenecks in the ER such as conflicting views on the values relevant for triage (see Chapter 6).

Commuting Routines by showing via simulations that large groups have a higher chance to cluster towards the same routine behaviour than small groups (see [217]) and that different habitual theories lead to different predictions regarding transport mode behaviour (see Chapter 7).

Animal Consumption Routines by enabling research on the connection of animal consumption routines with shopping and cooking routines and the interplay between animal consumption habits and the collective view on animal consumption (see [167] and Section 8.2.2).

More generally, SoPrA provides the theoretical groundwork that promotes policy modelling where policies are (1) comparable due to SoPrA's domain-independence, (2) trustworthy due to SoPrA's evidence-driven approach, (3) understandable due to SoPrA's parsimony (4) socially supported due to the SoPrA's ability to capture both individual and social values.

Appendix A: Overviews

Statement of Ethics

All data gathered is in accordance with Dutch law for ethical data collection. All the code and data are freely available to enable reproduction of the presented results.

Simulation Models and Online Repositories

The dissertation 'Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models' used simulation models throughout the thesis. This document gives an overview of where these models are found. The models are stored in one place to avoid confusion about the most recent model. The models are stored in GitHub to enable fast-updating, version history and collaboration.

Social Practice Agent Framework and Applications

The Social Practice Agent framework is modelled in Repast and OWL. The framework is available as a template. Two specific models used based on earlier versions on SoPrA are used in the thesis.

Table 1: Overview of the online repositories for the SoPrA framework.

Title	Ch.	Link
Most Recent SoPra Framework (Repast, OWL)	4	https://github.com/rmercuur/SoPrA-Models
Older Version (OWL)	6	https://github.com/rmercuur/SOPRA
Older version (Repast)	8	https://github.com/rmercuur/HabitsTraffic

Ultimatum Game Simulation Models

Models that simulate the behaviour of a homo economicus agent, a value-based agent and a norm based agent in a psychological experiment called the ultimatum game.

Table 2: Overview of the online repositories for the Ultimatum Game simulation models.

Title	Ch.	Link
Most Recent UG Simulation (Repast)	5	https://github.com/rmercuur/moral-ultimatum-game
Older Version UG Simulation (Repast)	2	https://github.com/rmercuur/UltimateValuesEclipse
Value-Based Agent Model (R)	2	https://github.com/rmercuur/UltimatValuesR
ML Model (Matlab) and Data Analysis (R)	5	https://github.com/rmercuur/moral-ultimatum-game

SoPrA Versions

SoPrA has been implemented in Repast Simphony, and Protégé. The code is available at Github (see 8.3).

SoPrA Repast Repast Simphony is a java-based tool for programming ABS based on the Eclipse IDE [198]. The translation from the UML to java code is relatively straight-forward as both are based on the object-orientated (OO) design method. Our implementation uses the `Table` class from Google's Guava to simplify the classes in the `activityCAssociations` and `agentAssociation` package. For example, the `HabitualConnection` class is transformed into a table in the agent class that maps a `ContextElement` and `Activity` pair to three doubles: the habitual strength, the personal view on the habitual strength and the collective view on the habitual strength (e.g., `Table<ContextElement, Activity, StrengthValues<Double, Double, Double>`). Our implementation comes with an example simulation that uses SoPrA to simulate the dynamics of commuting behaviour based on Chapter 7. Researchers who aim to use SoPrA to simulate routines are recommended to use SoPrA Repast.

SoPrA OWL Our online repository provides a formal framework in the OWL-language. A formal framework promotes unambiguity, rigour and enables formal reasoning. Protégé is a tool to author formal ontologies and enable formal reasoners that aid the modeller [189]. Using Protege –a formal reasoner – we show the framework is consistent, can verify our domain models correctly implement the framework and verify properties. In detail, Protégé enables modellers to make domain-specific models and use the formal reasoner to check if these models corresponds to the domain-independent framework we made. Protégé uses description logic: a logic that trades its expressivity to acquire decidability [128]. The UML classes correspond to OWL classes and UML associations to OWL properties. Chapter 6 describes the translation process from (an earlier version of SoPrA) to description logic. This shows that description logic is expressive enough to capture the semantics of SoPrA and that the formalization of SoPrA in Protégé is satisfiable. Protégé can be used in combination with ABM to infer new knowledge for the agents. For example, one can use the formal reasoner to apply to automatically infer that if riding a bike, taking the train to work and walking promote environmentalism then non-car commuting also promotes environmentalism. In short, researchers who want to verify that their domain model corresponds to SoPrA or want to infer new domain knowledge using formal reasoning are recommended to use SoPrA OWL.

SoPrA Framework Extensions

Figure 1 provides an extension of our core SoPrA framework. We treat concepts that are mentioned in related work but are not necessary to fulfil our requirements.

Roles Roles enable modellers to group similar agents. Grouping agents has several advantages. First, instead of defining multiple agents that have the same beliefs (e.g., car driving is efficient), modellers can define a role (e.g., a parent) that believes car driving is efficient. Figure 1 reflects the possibility for the `Role` class to have the same connections as the `Agent` class with the generalization arrow. Connecting a belief once to a role instead of connecting multiple beliefs to multiple agents reduces the spatial complexity of the model. Second, roles enable agents to enact multiple roles and thus entertain mul-

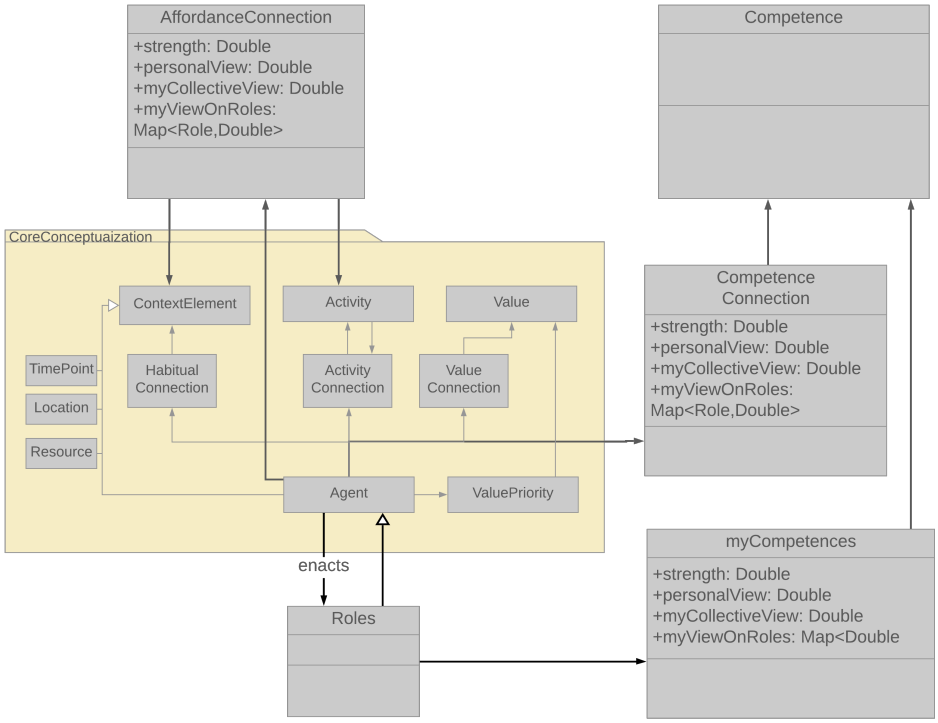


Figure 1: A UML diagram that extends the core SoPrA framework adding affordances, competences and roles.

tiple views. This is reflected in the `enact` association from the `Agent` class to the `Role` class. Third, roles enable an agent to hold a view on a specific group of agents. This feature is reflected in the attribute `myViewOnRoles` in the classes that connect agents and elements. This attribute maps a `Role` set (i.e., groups of agents) to a `Double` representing how strongly the agent believes this group of agents agrees on the connection (e.g., parents believe car driving is efficient, parents believe car-driving requires the competence to drive a car).[5]

Affordances Affordances specify for a context-element what actions it makes possible [102]. For example, a chair affords to sit. Affordances are related to habitual triggers. Affordances specify what actions are possible, habitual triggers which actions are salient [69]. Our extended framework contains the `AffordanceConnection` class as an activity association analogue to habitual triggers. Specifying the `strength` of the connection as a `Double` enables both a more restricted view on affordances (e.g., a chair affords sitting so the

[5] Although this attribute enables the modeller to specify beliefs about groups of agents for *any* connection (e.g., a habitual connection, a value connection), for ease of presentation Figure 1 only depicts this attribute in the classes related to the extensions .

strength is 1, but a table does not so the strength is 0) as well as a continuous view on affordances (e.g., a chair is better for sitting than a table).

Competences A competence is the capability to make a certain proposition true [76]. Competencies are criteria within the agent that enable an action (whereas affordances are criteria outside the agent that enable an action). In Shove et al. [235] competencies represent the implicit skill or know-how that is needed to act. In SoPrA-extended competences are added as both an activity association (i.e., `CompetenceConnection`) and an agent association (i.e., `myCompetences`) analogue to values. Specifying the `strength` of the connection as a double enables both a more restricted view on competences (e.g., a driver needs to know how to stay calm to drive) as well as a continuous view on competences (e.g., it is better when a driver knows how to stay calm when driving).

Requirements

Table 3 provides an overview on the requirements on habituality, sociality and interconnectivity.

Table 3: An overview on the requirements on habituality, sociality and interconnectivity

	We require the model to...		
Nr.	Habituality	Sociality	Interconnectivity
1	capture that social practices are habitual as they are similar over time	capture that social practices are social as they are similar over people.	capture that social practices are interconnected as they are similar over activities.
2	provide the primary concepts and relations to researchers to model habitual decisions, updates and reasoning	provide the primary concepts and relations to enable agents to make socially intelligent decisions, update social information and reason about collective concepts.	provide the primary concepts and relations to enable agents to make interconnected decisions, updates and reason about the interconnectedness of activities.
3	enable agents to differentiate between habits and intentions	use a comprehensive set of collective concepts that support social decision-making, updating and reasoning.	express that social practices are connected in terms of time, space and common elements.
4	support habitual relations between an action and a context-element where the strength of that relationship is a continuous parameter.	support agents that order social information around their practices.	differentiate between different types of activities: atomic activities and sequential activities (an ordered sequence of actions).
5	capture that context-elements can comprise resources, activities, location, timepoints or other people.	capture that social practices relate the collective social world with the individual world of interactions.	capture both temporal and ontological relations between activities to enable agents to make decisions and inferences.
6	to capture that agents can differ in the strength of a connection, maximum strength, time to reach this maximum, amount of attention they attribute to an action	capture that agents have a personal view on a social practice.	
7	to provide a concept that captures the abstract aim intentions are directed at.	capture that agents have a collective view on a social practice.	
		enable agents to have a different personal view than their collective view.	
		enable agents to each have a different collective view on a social practice.	

Appendix B: Towards an ABM of Gossip in Organizations

Introduction

The phenomenon of rumourmongering has malicious impacts on societies. Rumours make people nervous, create stress, shake financial markets and disrupt aid operations [259]. In organizations, rumours lead to unpleasant consequences such as, breaking the workplace harmony, reduction of profit, drain of productivity and damaging the reputation of a company [67, 181]. Recent work on the McDonald's wormburger rumour and the P&G Satan rumour confirm the negative impact of rumours on the productivity of firms [67].

For more than hundred years, scholars from a wide range of disciplines are trying to understand different dimensions of this phenomenon. Research in rumour studies can be classified according to the approach followed: a case-based approach and a model-based approach. In the case-based approach, results are based on case studies, not on models, making it hard to generalize their conclusions. The model-based approach tries to explain the phenomenon of rumours by model-based based simulations. The model-based approaches, so far, focus only on the dynamic of the spread, while rumour is a collective phenomenon and the acts of individuals can influence the whole system. Rumours in organizations have been mainly approached with case-based studied and dynamic spreading model. To our knowledge there are no studies where the cognition of the individual is taken into account.

In our agent-based approach, we study the dynamics of the spread of rumours in organizations as an emergent (collective) behavior resulting from the behavior of individual agents using social practice theory. We use the proposed model to study the impact of change in organizational layout on control of organizational rumour.

The concept of social practices stems from sociology, and aims to depict our 'doings and sayings' [227, p. 86], such as dining, commuting and rumourmongering. This chapter uses the semantics of the social practice agent (SoPrA) model [172] to gain insights in rumourmongering in organizations.[6] SoPrA provides an unique tool to combine habitual behavior, social intelligence and interconnected practices in one

This chapter has been published as A. E. Fard, R. Mercuur, V. Dignum, C. M. Jonker, and B. van de Walle, "Towards Agent-based Models of Rumours in Organizations: A Social Practice Theory Approach," in Advances in Social Simulation, 2020, pp. 141–153. [86]

[6] Mercuur et al. [172] provides a static model of SoPrA based on literature and argued modelling choices. This chapter applies this model to the domain and extends it by including competences and affordances and modelling a dynamic component based on [167]. Note that Mercuur et al. [172] is still under review and only available as pre-print at the moment of writing.

model. This makes SoPrA especially well-suited for studying the spread of rumours in organizations as this practice is largely habitual, social [97] and interconnected with practices as working and moving around. Given the lack of available empirical data on the social practice of rumourmongering, we give a proof-of-concept on how to collect data by doing eight semi-structured interviews.

This chapter is organized as follows. The next section provides an overview of the research on rumours with an emphasis on studies of organizational rumour. Section Domain describes the context for our experiment, and the methodology of data collection and data preprocessing. The model is introduced in Section Model. One possible experiment is described in Section Experiment and Section Conclusion presents our conclusions, discussion and ideas for future work.

Background & Related Work

Rumours are unverified propositions or allegations which are not accompanied by corroborative evidence [67]. Rumours take different forms such as exaggerations, fabrications, explanations [210], wishes and fears [149]. Rumours have a lifecycle and change over the time. Allport and Postman in their seminal work "psychology of rumour" concluded that, "as a rumour travels, it grows shorter, more concise, more easily grasped and told." [119]. Buckner considers rumour a collective behaviour which is becoming more or less accurate while being passed on as they are subjected to the individuals' interpretations which depends on the structure of the situation in which the rumour originates and spreads subsequently [40]. Rumours are conceived to be unpleasant phenomena that should be curtailed. Therefore, a number of strategies have been proposed to prevent and control them [40, 151, 200].

One of the rumour contexts that has received attention from researchers for almost four decades is organizations. Like rumour in general context which is explained in above paragraph, rumour in organizations has different types and follow its own life-cycle [31–33, 66, 67, 146]. Also, to quell credible and non-credible organizational rumours, a number of different techniques and strategies have been suggested [67, 146]. The research approach also follow the same pattern, with a slight difference which to best of our knowledge is qualitative without adopting any modelling approach.

The related literature reported above are based on case studies or experiments in the wild. This pertains to the types of rumour, dynamics of rumour and strategies to control rumours, either in general or in organizational contexts. These case studies and experiments are to inform the construction of theories and models underlying the phenomenon of rumourmongering. Theories and models, in turn, should be tested in case studies and simulations. Model-based approaches do just that. However, the current state-of-the-art in model-based simulations of rumourmongering focus only on the dynamics of the rumourmongering, comparable to the epidemic modelling and spread of viruses [55, 193, 251, 262, 272]. These models do not consider the complexities of the agents that participate in rumourmongering.

The research area of agent-based social simulations (ABSS) specializes on simulating the social phenomena as phenomena that emerge from the behaviour of individual agents. ABSS is a powerful tool for empirical research. It offers a natu-

ral environment for the study of connectionist phenomena in social science. This approach permits one to study how individual behaviour give rise to macroscopic phenomenon [82]. Such an approach is an ideal way to study the macro effects of various social practices, because it can capture routines which are practiced by individuals on a regular basis in micro level and see their collective influence in a macro level.

Domain

This research investigates the daily routine of rumourmongering in a faculty building on the campus of a Dutch University. In this faculty, students, researchers and staff work in offices with capacity of one to ten people. Aside from the actual work going on in the building, filling a bottle with water, getting coffee from the coffee machine, having lunch at the canteen and going to the toilet are among the most obvious practices that every employee in this faculty does on a daily basis.

Nevertheless, there are other daily routines in the organization which are not that obvious. One of these latent routines is rumourmongering. Rumours or unverified information are transferred between students, researchers and staffs on a daily basis, during lunch, while queuing for coffee, when seeing each other in the hallways, and when meeting in classrooms and offices. All these situations are potential contexts for casual talks and information communication without solid evidence.

For data collection we conducted semi-structured explorative interviews with people from the above-mentioned faculty. Semi-structured interviews allows us to ask questions that are specifically aimed at acquiring the content needed for the SoPrA model, while still giving the freedom to ask follow up questions on unclear answers. The data collection can be improved in future works by increasing the number of interviewees and diversifying them (Not only asking from students). For demographic information, the reader is referred to Table 4. We prepared following question set to ask from each interviewee based on the meta-model which will be explained in the next section:

1. What are the essential competencies for rumourmongering?
2. What are the associated values with rumourmongering?
3. What kind of physical setting is associated with rumourmongering?

Table 4: The demography of the interviewed subjects.

Number of Interviews	Number of Different Countries	Lowest Educational Level	Mean age	Female %
8	6	MSc	28	50

Given the thin line between personality traits and competences, we used the Big Five model [110] to differentiate between personality traits and competences. For Question 2, we asked the interviewees to choose the relevant values from Schwartz's Basic Human Values model [229]. We asked the same set of questions about fact-based talk.

We processed the collected data in two ways before using it in the model. Firstly, we clustered answers that point to the same concept. For example, in Question 3, interviewees gave answers such as cafeteria, coffee shop and cafe to point to a place where people can get together and drink coffee. In the coffee example, we clustered answers under the term of "coffee place".

Secondly, we classified the answers to Question 2. As mentioned, for that question, we asked interviewees to pick associated values from Schwartz's Basic Human Values model. We used the third abstraction level of the model which is more fine-grained and compared to other levels, and gave the interviewees a better idea of what they point to. However, a model based on level three, would not allow us to compare the agents effectively. Therefore, we decided to wrap the answers and classify them based on second abstraction level. Using a classification based on the first abstraction level would have been too homogeneous in the sense that the agents would behave too similar, which would loose the effectiveness of the simulation.

Model

The model has two main parts: (i) static part and (ii) dynamic part. In the static part, the components of the model and their properties are described, and in the dynamic part we explain the interaction of those components.

Static Part

This section describes the SoPrA meta-model which is used as the groundwork for our agent-based model, how we use empirical data to initiate the model, the model choices we make and how we tailor the model to the context of organization.

The SoPrA meta-model was introduced by Mercuur et al. [172] and describes how the macro concept of social practices can be connected to micro level agent concepts. Figure 2 shows SoPrA in a UML-diagram. The main objects in a SoPrA model are activities (e.g., fact talk, rumourmongering), agents (e.g., PhD students, supervisors), competences (e.g., networking, listening), context elements (e.g., office, cafetaria) and values. Values here refer to human values as found by the earlier stated Schwartz model, such as, power or conformity. The social practice is an interconnection of (1) activities and (2) related associations as depicted by the grey box in Figure 2. For example, the practice of talking consists of two possible activities fact talk or rumourmongering. The social practice connects these different activities with the `Implementation` association. If activity A implements activity B this means that A is a way of or a part of doing B.

The `Implementation` association is the first of several associations that are related to an activity (see Table 5). Most associations are fairly self-explanatory,

however the `Trigger` and `Convention` attribute are a bit more complex. Following Wood and Neal [267], triggers are the basis for habitual behaviour. If an agent is near a context element that has a trigger association with an activity, then it will do that activity automatically (without for example considering its values). Following Crawford and Ostrom [53], conventions are related to norms and signify that something is the normal way to do something.. If an agent believes that activity A is a strategy for activity B, then it believes that other agents usually implement activity B by doing activity A.

The SoPrA meta-model does not only relate the activities to other classes, but the agent itself also has two types of associations: `HasCompetence` and `ValueAdherence` which plays a role in choosing the activities it will do:. The `HasCompetence` association links possible skills to the agent who masters those. The `ValueAdherence` association captures if an agent finds that value important.

Table 5: The associations attached to the activity and their specification.

Association	**Specification**
Implementation	which activities are a way of or a part of doing the activity
Affordance	which context elements are needed to do the activity
RequiredCompetence	which competences are needed to do the activity
Knowledge	which activities an agent knows about
Belief	which personal beliefs an agent has about the activity
RelatedValue	which values are promoted or demoted by the activity
Trigger	which context elements habitually start the activity
Convention	which activities usually implement the activity

The model can be initiated using empirical data. Note that in this study we focussed on a small set of explorative interviews. We show with this initial data a proof-of-concept of how the model can be initiated. To properly ground the model a larger and more rigorous empirical study is necessary.

The activity class has three instances: talking, rumourmongering and fact talk. The number of instances of agent can vary in the different experiments (see Section 8.3). The instances of the context element, competence and values class are based on the gathered data and can be found in Table 6 and 7.[7] The complete static model consists both of object instances and associations between these instances. An example focusing on one agent (i.e., `Bob`) and one activity (i.e., `rumourmongering`) is shown in Figure 3. Bob believes that the activity of rumourmongering is related to the value of privacy, curiosity and social power. He thinks it requires the competence of networking and noticing juicy details and thinks the activity is triggered (to some extent) by the hallway, restaurant and another agent named Alice.

[7] The context-element 'Friend' and 'Colleague' are special cases; these are rather attributes of context-elements (i.e., agents) than context-elements themselves. In our model these are to some extent implicitly captured, because the agents who one sees most often (i.e., friends, colleagues) are mostly likely to be habitually associated with an action.

Furthermore, he himself has the competence of networking and adheres strongest to the value of ambition and weakest to the value of pleasure.

Table 6: The elements associated with the rumourmongering activity.

Rumourmongering		
Context Elements	**Meaning**	**Competence**
Friend	Self-Direction	Sneaky Skills
Coffee place	Power	Network Skills
Hallway	Hedonism	Talking Skills
Restaurant	Achievement	Observing Skills
Office	Benevolence	
Phone		
Computer		

Table 7: The elements associated with the fact talk activity.

Fact talk		
Context Elements	**Meaning**	**Competence**
Colleague	Universalism	Being knowledgeable
Academic Staff	Self-Direction	Listening Skills
Office	Benevolence	Critical Thinking Skills
Conference	Achievement	Communication Skills
Meeting room	Tradition	
Classroom		
Restaurant		
Phone		
Computer		
Pen		
Coffee		

The agents differ in which activity they associate with which element. In other words, the SoPrA meta-model does not initiate one social practice that all agents share, but one social practice *per agent*. The chance that an agent relates an activity to a competence is based on the empirical data we gathered in the interviews. For example, if 50% of the interviewees linked critical thinking skills to fact talk the chance an agent makes this association depends on a binomial distribution with $p = 0.5$. For `relatedValue` association and `HabitualTrigger` association all agents make the associations as mentioned in Table 6 and 7. However, the weights differ per agent. The weights for the `relatedValue` association are picked from a normal distribution between 0 and 1. Given the lack of empirical data on the relation between activities and human values, we follow the related finding of the World Value Survey that people adhere to values with roughly a normal distribution [4]. The weights for `HabitualTrigger` are picked on a logarithmic distribution based on the empirical work of [155]. One interesting modelling choice we made was to drop the `Affordance` assocations in the conceptual model. The SoPrA meta-model conceptualizes two associations with context elements. The `HabitualTrigger` association representing that some context element can automatically lead to a reactive action and the `Affordance` association representing that some context elements are a pre-condition to enact a certain behaviour. None of our interviewees

mentioned a possible context element that affords rumourmongering fact talk. As such this association seemed irrelevant for our model.

The associations related to the agents themselves are based on random distributions. Each competence has a 50% chance to be related to an agent. Each value is associated to each agent, but the weights differ. The weights for the `hasValue` association strength is based on a correlated normal distribution. Schwartz [229] shows that the strength to which people adhere to values is correlated. For example, people who positively value universalism usually negatively value achievement. We use the correlations found by Schwartz [229] to simulate intercorrelated normal distribution from which we pick the weights. In future work, we aim to extent our interviews to also gather data that can inform these weights.

For our modelling context, we need to extend the SoPrA model with a spatial component. We do this by adding two attributes to the `ContextElement` class called `x-coordinate` and `y-coordinate`. These coordinates can be used by the agent to sense which objects are near. Note that every agent is also a context element as indicated with the 'generalization' association in the UML-diagram.

Dynamic Part

This section describes the dynamic part of the model which on each tick comprises:

1. An agent decides on its location using the moving submodel and updates its coordinate attributes.
2. An agent decides if it will engage in fact talk or rumourmongering based on the choose-activity submodel.

The moving submodel has four components that agents can transfer between. As it is shown in Figure 4 the initial state is offices and from that state agents can leave their offices and pass the hallway to either have lunch at the restaurant or grab a cup of coffee at the coffee place. During the interviews, we discovered most of the people do those daily routines around the same period of time and only a few people do not follow this pattern and leave their offices out of usual time periods, so we concluded the transition of agents between different locations is a random phenomenon which follows a normal probability distribution.

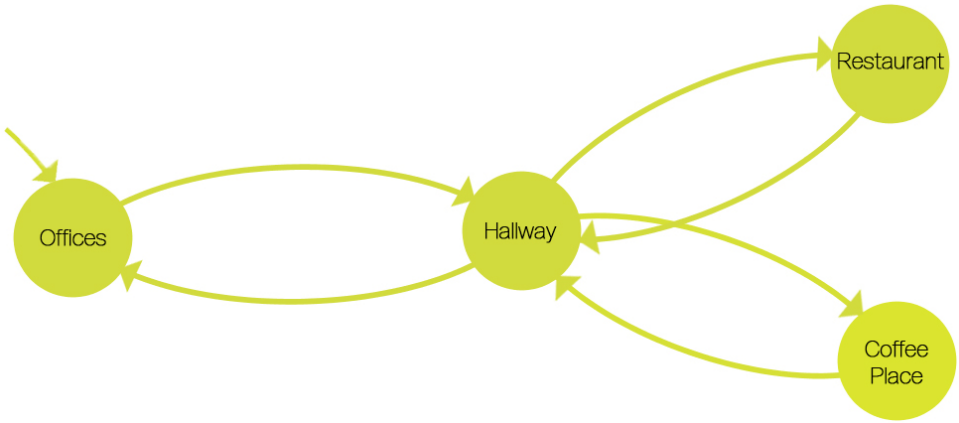


Figure 4: The moving model for agents

The choose-activity submodel is based on Mercuur et al. [169] and has three stages. The submodel is depicted in Figure 5. The agent starts by considering both rumourmongering and fact talk. At each stage the agent makes a decision on one cognitive aspect. If this aspect is not conclusive it will prolong the decision to the next stage. In the first stage, the agent compares its own competences to the competences that it beliefs to be required for the activity. In our example model depicted in Figure 3, Bob would decide it cannot do the activity of `rumourmongering`, because it requires a competence he does not have: noticing juicy details. As such, Bob will engage in fact talk. (Note that if Bob does not have the skill to do either activity, then the decision is also prolonged to the next stage.) In the second stage, an agent tries to make a decision based on its habits. It will survey its context and decide which context elements are near, i.e., resources, places or other agents. If it has a habitual trigger association with a particular strong strength between one of those context elements and either rumourmongering or fact talk it will automatically do that action. In the last stage, the agent will consider how strongly it relates certain values to both activities and how strongly it adheres itself to these values. Consequently, it makes a comparison between the two activities and decides which best suits its values. For the complete implementation of the habitual model and value model we refer to [167].

Experiment

The proposed rumour model with elements associated with physical settings, individuals' values and competencies enables us to investigate impacts of a variation of settings and interventions on the spread of rumours in organizations.

One of the open questions in organizational rumour literature is the effectiveness of different prevention and control strategies. In our approach we only need to extend the model with the specific elements and characteristics of the case that we would like to study. In this chapter we study the effect of organizational layout on rumour dynamics. In our case, we take the size of offices and number of coffee

places as the proxies for organizational layout and juxtapose two organizational layouts cases (Figure 4) to understand the impact of layout on rumourmongering dynamic.

To setup the model, we determine the number of agents, then initialize the context and agents. In the organization that we studied each section has on average 50 people, therefore, we pick 50 as the number of the agents. For context initialization, we design the layouts and assign agents to different locations, then we initialize agents with probability distributions for routines such as grab a cup of coffee or having lunch. After the model setup, it can be executed.

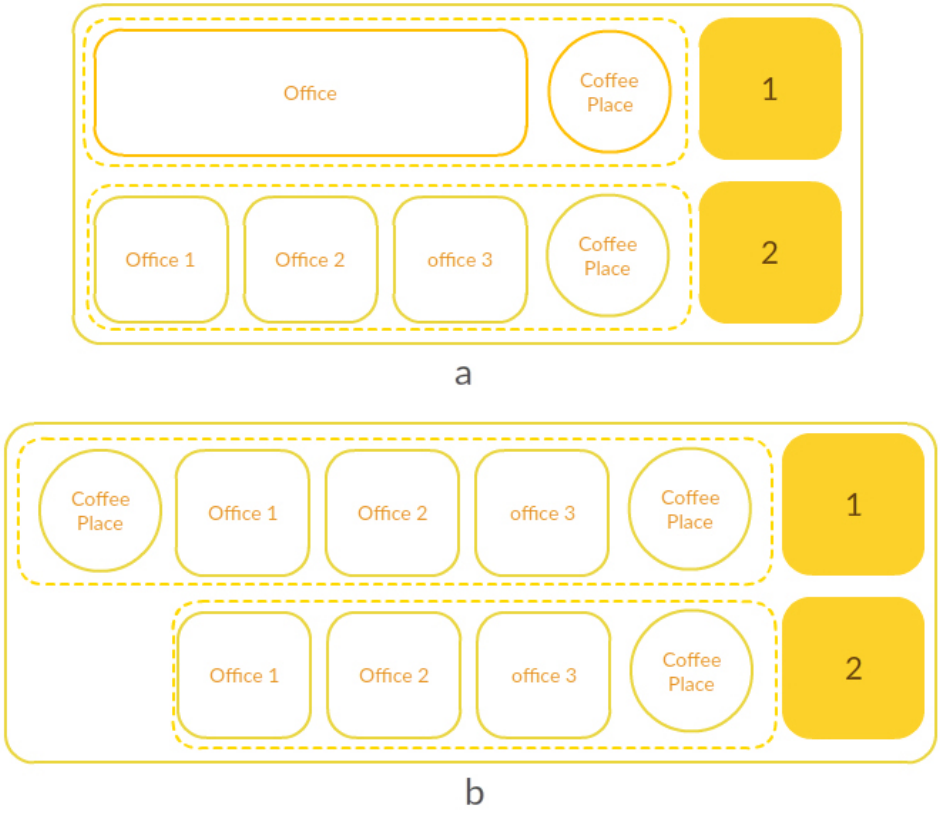


Figure 6: (a) In this case, we study the impact of office size on dynamics of rumourmongering (b) In this case we study the impact of number of coffee places on the dynamics of rumourmongering

Discussion & Future Research

Modelling rumourmongering has been studied since 1964. So far, the modelling did not consider the complexities of individual agents, and mostly focused on the spreading behaviour of the phenomenon. In the model proposed in this chapter,

agents have a cognitive layer that deploys social practice theory and views rumour as a routine with associated competencies, values and a physical setting.

In this research, we narrowed our study to the context of organization and after introducing the generic model, we tailored our model to the context of organization via empirical data collected though interviews conducted in a Dutch University. Based on explorative interviews we established that social practice theory are likely to be applicable as people shared a view on rumour, and their habits regarding rumour and rumours seem to be intertwined with other activities.

Our model can be used to study a wide range of topics in organizational rumour studies, in particular for testing the effectiveness of interventions for prevention and control of rumours in organizations.

Future work is to extend the questionnaire by asking about associations, conduct more and more rigorous interviews, implement the model and run the proposed experiments that explore different organization layouts. Furthermore, we aim to validate our model by looking at how rumours travel from person to person in the organization during a pre-selected time period.

Contributions & Acknowledgements

Ebrahimi Fard & Mercuur wrote the first draft. Ebrahimi Fard provided the domain knowledge and collected most data, whereas Mercuur provided the meta-model and methodological knowledge. Dignum, Jonker and van der Walle. supervised the process and contributed to the draft by providing comments, feedback and rewriting. This research was supported by the Engineering Social Technologies for a Responsible Digital Future project at TU Delft and ETH Zurich.

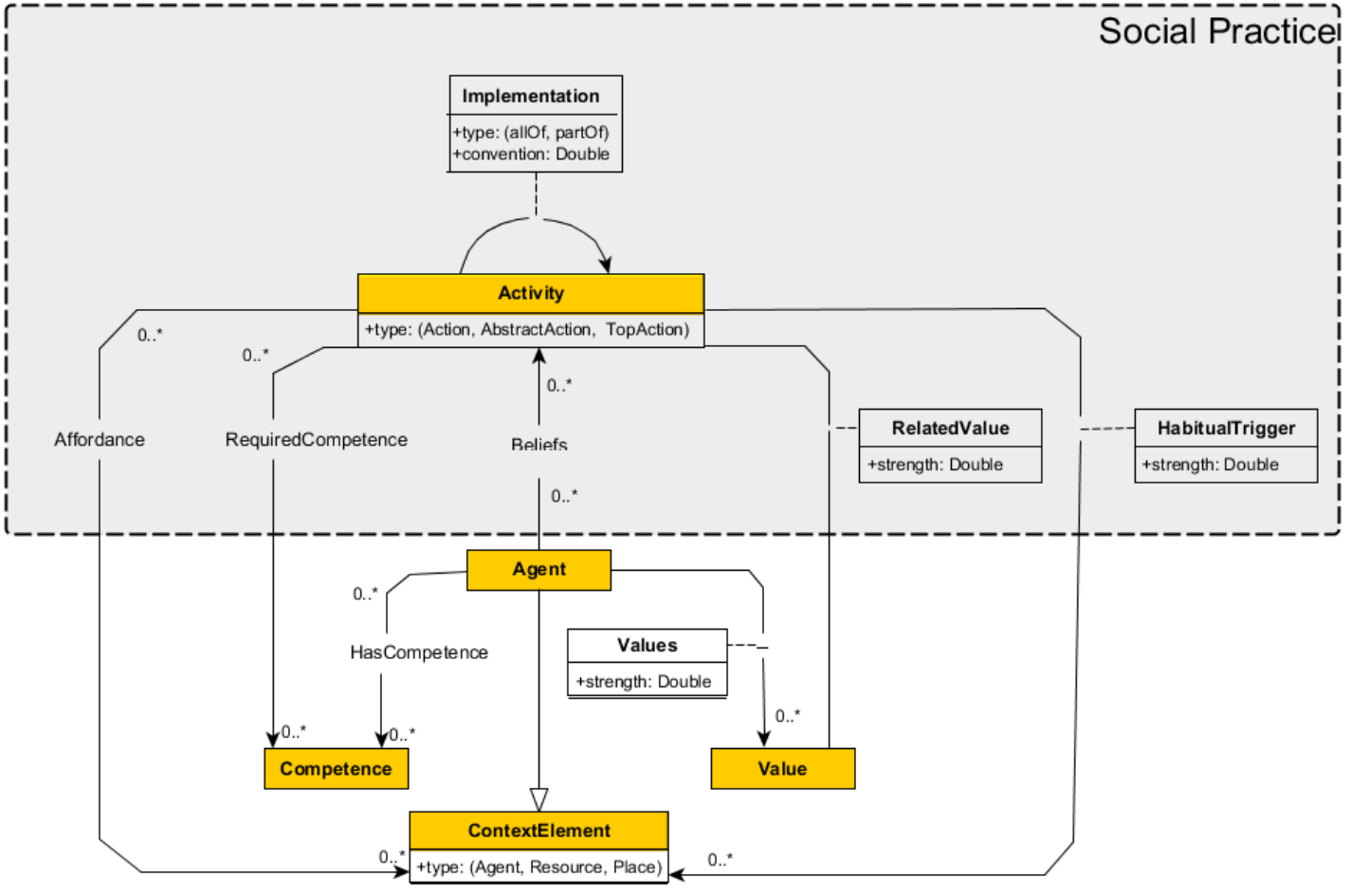


Figure 2: The social practice meta-model captured in the Unified Modelling Language, including classes (yellow boxes), associations (lines), association classes (transparent boxes), navigability (arrow-ends) and multiplicity (numbers).

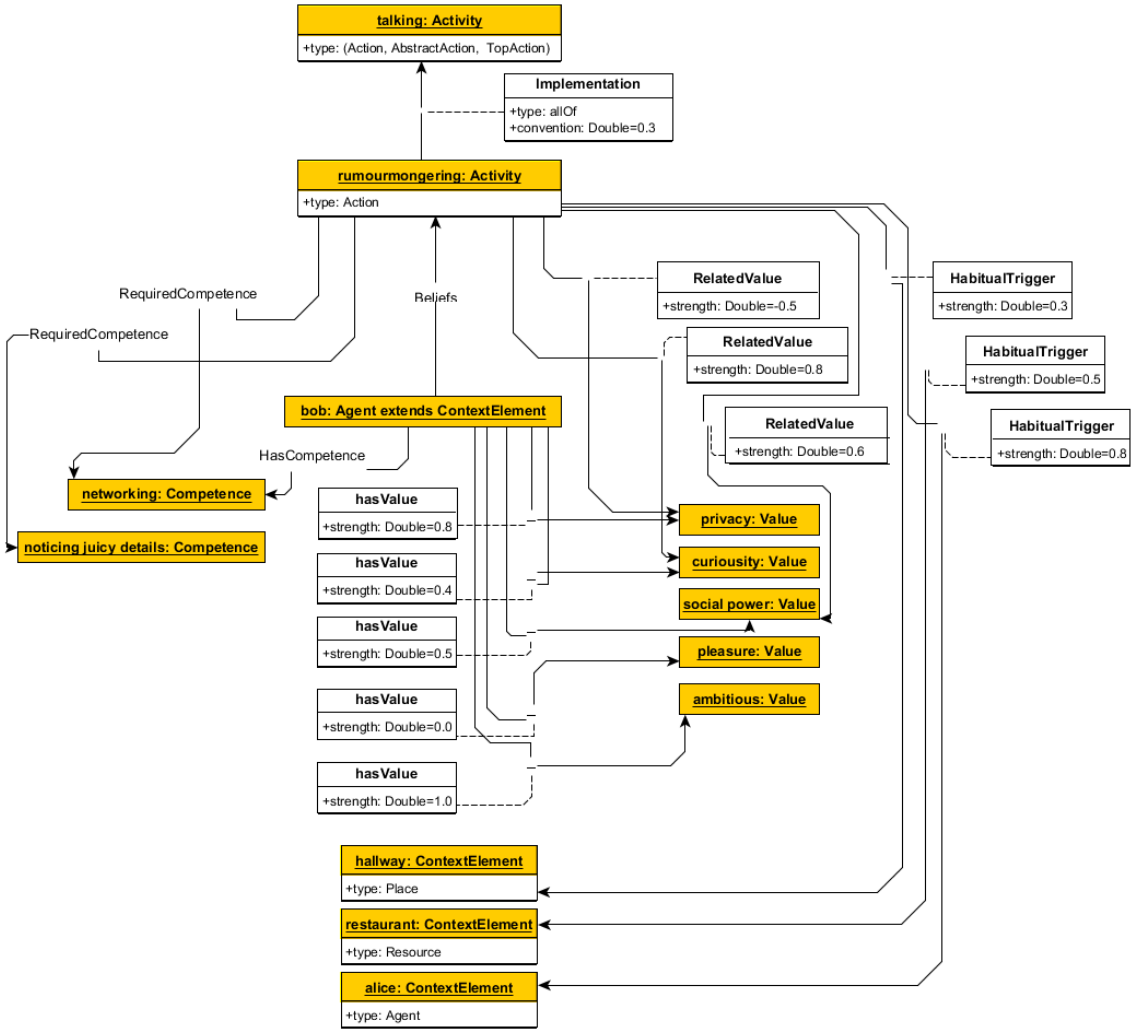


Figure 3: An instance of the SoPrA meta-model for the activity of rumourmongering and one agent. For illustration purposes the assocations related to the activity 'talking and the agent 'alice' are omitted.

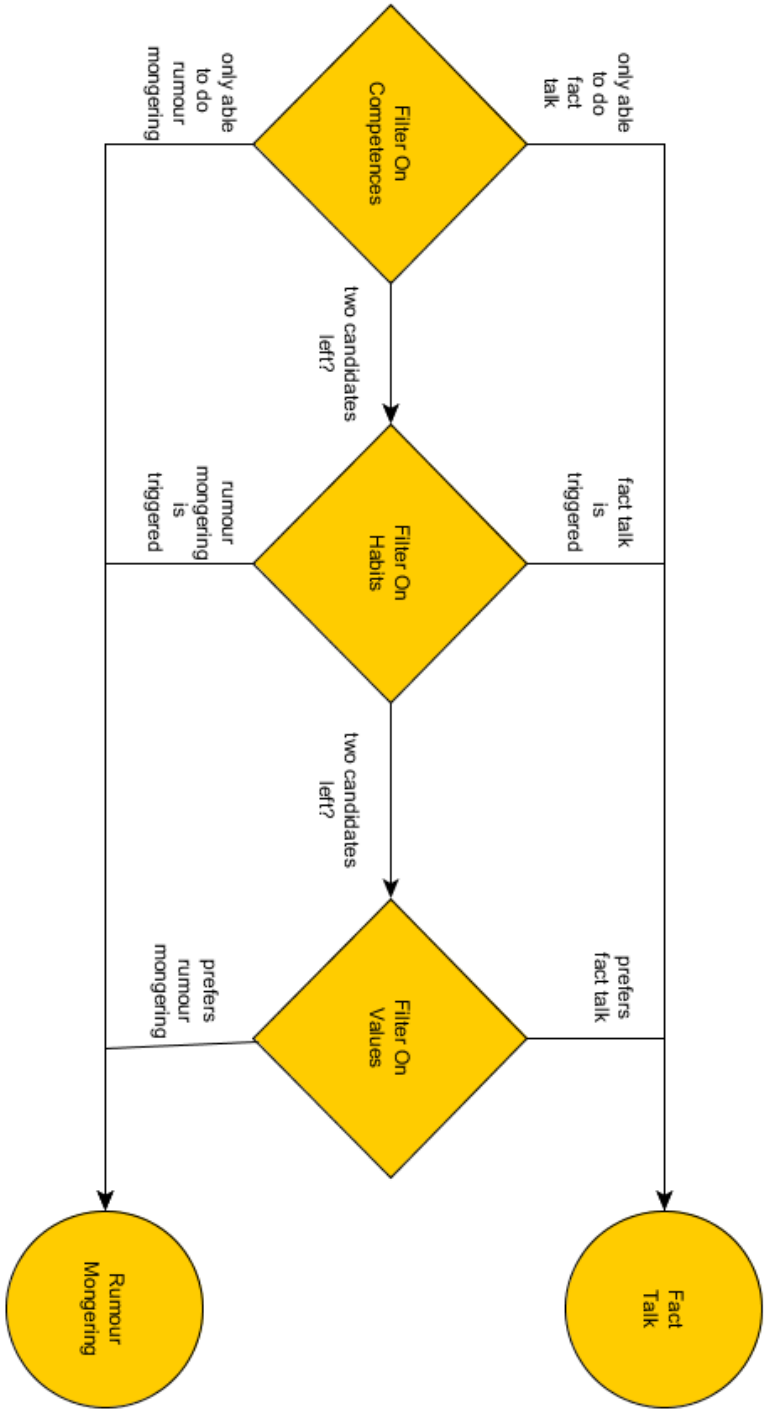


Figure 5: The choose-activity submodel and the three stages the agent uses to decide on its activity: competences, habits and values.

SIKS Dissertation Series

2011 01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models

02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems

04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference

05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage

07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction

08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues

09 Tim de Jong (OU), Contextualised Mobile Media for Learning

10 Bart Bogaert (UvT), Cloud Content Contention

11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective

12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining

13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling

14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets

15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval

16 Maarten Schadd (UM), Selective Search in Games of Different Complexity

17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness

18 Mark Ponsen (UM), Strategic Decision-Making in complex games

19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles

20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems

22 Junte Zhang (UVA), System Evaluation of Archival Description and Access

23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media

24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics

26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

29 Faisal Kamiran (TUE), Discrimination-aware Classification

30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science

33 Tom van der Weide (UU), Arguing to Motivate Decisions

34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training

36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization

39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

44 Boris Reuderink (UT), Robust Brain-Computer Interfaces

45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression

48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

---

2012 01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications

05 Marijn Plomp (UU), Maturing Interorganisational Information Systems

06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks

07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories

09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms

10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment

11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems

13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems

15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.

16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment

17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance

18 Eltjo Poort (VU), Improving Solution Architecting Practices

19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution

20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval

22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction

24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval

25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application

26 Emile de Maat (UVA), Making Sense of Legal Text

27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games

28 Nancy Pascall (UvT), Engendering Technology Empowering Women

29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval

30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making

31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure

32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning

33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)

34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications

35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics

36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes

37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation

38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms

39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks

40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia

41 Sebastian Kelle (OU), Game Design Patterns for Learning

42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning

43 Withdrawn

44 Anna Tordai (VU), On Combining Alignment Techniques

45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions

46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation

47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior

48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data

49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions

50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering

51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

---

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support

02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing

03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics

04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling

05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns

06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience

07 Giel van Lankveld (UvT), Quantifying Individual Player Differences

08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators

09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications

10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.

11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services

12 Marian Razavian (VU), Knowledge-driven Migration to Services

13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly

14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning

15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications

16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation

17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid

18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification

19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling

20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval

21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation

22 Tom Claassen (RUN), Causal Discovery and Logic

23 Patricio de Alencar Silva (UvT), Value Activity Monitoring

24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning

25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System

26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning

27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance

28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience

29 Iwan de Kok (UT), Listening Heads

30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support

31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications

32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development

33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere

34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search

35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction

36 Than Lam Hoang (TUe), Pattern Mining in Data Streams

37 Dirk Börner (OUN), Ambient Learning Displays

38 Eelco den Heijer (VU), Autonomous Evolutionary Art

39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems

40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games

41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning

42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning

43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

---

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data

02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method

03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions

04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation

05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability

06 Damian Tamburri (VU), Supporting Networked Software Development

07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior

08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints

09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language

10 Ivan Salvador Razo Zapata (VU), Service Value Networks

11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support

12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control

13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

14 Yangyang Shi (TUD), Language Models With Meta-information

15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare

16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria

17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability

18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations

19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support

20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link

21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments

22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training

23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era

24 Davide Ceolin (VU), Trusting Semi-structured Web Data

25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction

26 Tim Baarslag (TUD), What to Bid and When to Stop

27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty

28 Anna Chmielowiec (VU), Decentralized k-Clique Matching

29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software

30 Peter de Cock (UvT), Anticipating Criminal Behaviour

31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support

32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data

33 Tesfa Tegegne (RUN), Service Discovery in eHealth

34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.

35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach

36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models

37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying

38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.

39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital

40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education

41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text

42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models

43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments

44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.

45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach

46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity

47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

---

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response

02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls

03 Twan van Laarhoven (RUN), Machine learning for network data

04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments

05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding

06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes

07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis

08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions

09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems

10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning

11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins

12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks

13 Giuseppe Procaccianti (VU), Energy-Efficient Software

14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations

15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation

16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork

17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs

18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories

19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners

20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination

21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning

22 Zhemin Zhu (UT), Co-occurrence Rate Networks

23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage

24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation

25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection

26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure

27 Sándor Héman (CWI), Updating compressed colomn stores

28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO

29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains

30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning

31 Yakup Koç (TUD), On the robustness of Power Grids

32 Jerome Gard (UL), Corporate Venture Management in SMEs

33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources

34 Victor de Graaf (UT), Gesocial Recommender Systems

35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

---

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VU), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UVA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web

19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UVA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UVA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (UvT), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design

39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UVA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TUE), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UVA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VU), Logics for causal inference under uncertainty

23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VU), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (UvT), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TUE), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TUE), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TUE), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech

30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data

06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VU), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming

36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TUE), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models

12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15	Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17	Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20	Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21	Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24	Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25	Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28	Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29	Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30	Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31	Gongjin Lan (VU), Learning better – From Baby to Better
32	Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33	Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
34	Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
2021 01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

Curriculum Vitæ

I was born and raised in the Netherlands (1991). I believe in compassion and wisdom. I am curious about human behaviour. In particular, the telepathy of social intelligence that coordinates our actions. I believe in what's important over what's normal. My personality type is a campaigner. I play the guitar, in particular, blues guitar. I meditate and find joy and purpose in a Buddhist view on the world. I support the effective altruism movement. I am vegan as I believe its the most effective way to reduce suffering in this world. I cherish an open-mind and being open to change.

In 2009 I moved to Utrecht to study artificial intelligence at Utrecht University. I received my bachelor's degree in 2012 and received my master's degree (cum laude) in 2015. During my studies, I enjoyed helping out in several courses as a teaching assistant, travelling Asia for half a year and working on my master's thesis at the University of Melbourne, Australia. My master's thesis used simulations to compare interventions that promote vegetarian dining. For this work, I received the Best Thesis Award of the Graduate School of Natural Sciences in 2015.

This dissertation marks the end of my PhD at Delft University of Technology, The Netherlands. In my PhD, I enjoyed being part of the European Social Simulation Association (ESSA) by reviewing papers, presenting my work and organizing workshops. In my PhD, I created the SoPrA framework that enables simulation studies of human routines. I used these simulations to gain insights into social systems: environmental commuting, environmental dining, the spread of rumours, and staff and patient management at hospitals. I hope my PhD will establish an evidence-driven and positive view on humans and will help to gain insights that improve society, and in particular, counter climate change.

List of Publications

White Paper R. Mercuur. Realistic Agents with Social Practices. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1752–1754. International Foundation for Autonomous Agents and Multiagent Systems, 2018

Values and Norms R. Mercuur, V. Dignum, C. Jonker, and others. The value of values and norms in social simulation. *Journal of Artificial Societies and Social Simulation*, 22(1): 1–9, 2019

- **Preliminary study on values** R. Mercuur, V. Dignum, and C. Jonker. Modeling Social Preferences with Values. In *The Proceedings of the 3nd International Workshop on Smart Simulation and Modelling for Complex Systems @ IJCAI*, pages 17–28, 2017
- **Preliminary study on adding norms** R. Mercuur, V. Dignum, and C. Jonker. Using Values and Norms to Model Realistic Social Agents. In *29th Benelux Conference on Artificial Intelligence*, page 357, 2017

Literature Review R. Mercuur, V. Dignum, and C. Jonker. Integrating Social Practice Theory in Agent-Based Models: A Review of Theories and Agents. *IEEE Transactions on Computational Social Systems*, 2020

Conceptualising SoPrA R. Mercuur, V. Dignum, and C. Jonker. Modelling human routines: Conceptualising social practice theory for agent-based simulation. *arXiv*, page (under review), 2020. ISSN 23318422

Applications

- **Emergency Room** R. Mercuur, J. B. Larsen, and V. Dignum. Modelling the social practices of an emergency room to ensure staff and patient wellbeing. *Belgian/Netherlands Artificial Intelligence Conference*, pages 133–147, 2018. ISSN 15687805
- **Gossip** A. E. Fard, R. Mercuur, V. Dignum, C. M. Jonker, and B. v. d. Walle. Towards Agent-based Models of Rumours in Organizations: A Social Practice Theory Approach. In *Advances in Social Simulation*, pages 141–153, Cham, 2020. Springer
- **Value-Alignment** L. C. Siebert, R. Mercuur, V. Dignum, J. van den Hoven, and C. Jonker. Improving Confidence in the Estimation of Values and Norms. In A. Aler Tubella, S. Cranefield, C. Frantz, F. Meneguzzi, and W. Vasconcelos, editors, *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIII*, pages 98–113, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72376-7
- **Discerning Habits** R. Mercuur, V. Dignum, and C. M. Jonker. Discerning Between Two Theories on Habits with Agent-based Simulation. In N. Eggert, editor, *Proceedings of the Social Simulation Conference 2019*, page (in press), Mainz, 2019. Springer LNCS

Supervision on Master Thesis

Adoption of HV Panels J. D'Haens. *Yes to agreements, yes to interventions, yes to solar panels*. PhD thesis, TU Delft, 2018. URL http://resolver.tudelft.nl/uuid:c0876137-7949-4fa8-92c6-b772965a2351

Transport Mode Choice N. Roes. *Belief Dynamis Driven By Social Influence: A Social Practice-Based Model Of Belief Adaption Driven By Intragroup Interaction With Application to Transport Mode Choices*. PhD thesis, Delft University of Technology, 2019. URL http://resolver.tudelft.nl/uuid:bfdc3a61-62a4-43b4-95f3-abee2a06e945

Summary

Motivation The behaviour of groups of people plays an important but complex role in a wide range of social challenges such as disease outbreaks, climate change and coördinating a hospital. For example, in a disease outbreak, it is important to know where people go and to what extent they conform to social distancing. However, this is not easy to find out: individuals act differently, influence each other and adapt their behaviour over time. Agent-based social simulation (ABSS) enables researchers to simulate these complex aspects of human behaviour by directly representing individual entities (called agents) and their interactions [107]. ABSS studies are used to guide policies, expose theories, explain social phenomena, describe social systems, or illustrate key dependencies [107]. In short, ABSS studies provide and communicate insights on complex social systems.

Current frameworks for agent models do not support researchers in ABSS studies on human routines (see Chapter 3). These frameworks focus on reasoning aspects, social aspects or psychological mechanisms such as fear or needs. These aspects do not suffice to capture the key aspects of human routines: humans make habitual decisions, interconnect these habits throughout the day and use these interconnected habits as a blueprint for social interaction (see Chapter 3). Evidence from the social sciences shows the influence of routines on human decision-making (see Chapter 3). Empirically, meta-studies in transport choices [124], food choices [215] or recycling [148] consistently show that measures of habits (one aspect of routines) have a moderate to strong correlation with behaviour. Furthermore, general theories on human decision-making emphasize the key role of routines in behaviour [8, 137, 250]. To quote William James on habits: "Ninety-nine hundredths or, possibly, nine hundred and ninety-nine thousandths of our activity is purely automatic and habitual, from our rising in the morning to our lying down each night."[Ch. VIII] [136]. In addition to this evidence for the importance of the habitual aspect of routines, there is similar theoretical and empirical evidence for the importance of the social and planning aspects of routines (see Chapter 3). An agent framework that integrates human routines would enable ABSS researchers to gain new insights into social systems grounded on this theoretical and empirical evidence.

Purpose This thesis provides the SoPrA framework (i.e., the Social Practice Agent framework) that fulfills the following research objective:

Research Objective

Create a domain-independent agent framework that integrates theories on social practices to support the simulation of human routines.

This thesis uses social practice theory (SPT) to conceptualize an agent framework that integrates our human routines. SPT is a socio-cognitive theory applicable to model human routines as the theory aims to describe our 'daily doings and sayings' [227]. By building SoPrA upon a theory from sociology (and expending and refining the theory using related evidence), the framework fulfills two criteria that support the success of ABSS studies:

Grounding in evidence ensures we strengthen the mapping evidence on human decision-making into our model of human decision-making: agents. Well-grounded agent models produce well-grounded ABSS studies.

Reuseability ensures the models based on the framework are usable by different scientists, comparable with other models and are easy to extend or combine with other models. The reusability of a model is important to the ABSS field as a collective: reuseability promotes efficiency and allows models to evolve (i.e., select, reproduce, mutate) toward increasingly successful representations of humans. To ensure the reusability of our agents we need a domain-independent agent *framework* that enables ABSS researchers to create multiple domain-dependent agent *models*.

Outline & Results Part I contains one chapter which expands on our motivation, research objective and questions mentioned above. Part II of the thesis identifies the aspects of social practice theory that are relevant for agent-based simulation, distills requirements from the literature, reviews current agent frameworks and provides the SoPrA framework that satisfies said requirements.

Chapter 2 provides a preliminary simulation study that focusses on two aspects of routines: values and norms. We found that human values provide a universal, moral, operationalised, empirically grounded and a trans-contextual basis for intentional behaviour and that norms provide the dynamic social component for behaviour. Values and norms provide the necessary submodels for SoPrA and gives insight into what social practices (SPs) capture over and above values and norms.

Chapter 3 zooms out again and considers the SP as a whole. We review the literature on sociology, psychology and agent theory related to routines. In particular:

We find that to integrate ABSSs, agent theory and SPT, we need to focus on three key aspects of behaviour: the habitual, social and the interconnected aspects (this answers **RQ1**). These aspects reflect current evidence that people's decisions are automatic and contra-intentional, based on human values, based on a collective view on the world and depend on hierarchies of activities.

We provide a list of requirements for an agent framework that integrates SPT to enable simulations of human routines (this answers **RQ2**). These requirements provide a a clear relationship between SoPrA and current

empirical and theoretical evidence and ensure SoPrA provides a transparent anchor in further investigation.

We show that current agent frameworks do not support researchers in simulating human routines (this answers **RQ3**). In particular, current agent framework lack (1) support for context-dependent habits, individual learning concerning habits and explicit reasoning about habits; (2) a comprehensive set of collective concepts, an efficient ordering of social information around actions and an a connection between individual and collective concepts; and (3) explicit relations between activities, between each activity and each other model concept and hierarchies of activities.

Chapter 4 presents the Social Practice Agent (SoPrA) framework that integrates SPT and fulfils the requirements set out in Chapter 3 (this answers **RQ4**).

Part III describes three applications of SoPrA. These case studies show how our framework is able to support research in different domains with different purposes by using different parts of SoPrA (this answers **RQ5**).

Chapter 5 presents a study that focuses on the value-alignment problem: a well-known problem in AI where one aims to match the behaviour of an autonomous agent with human values. The study analyses to what extent an autonomous agent is able to estimate the values and norms of a simulated human.

Chapter 6 presents a study on identifying social bottlenecks in hospitals. The study provides a formal model (an OWL ontology) of the social dimension of the emergency room (ER) and uses a formal reasoner to find said bottlenecks.

Chapter 7 presents a study comparing two theories on habits via simulation. Via simulation, we identify an empirical experiment that enables social scientists to find out which theory is supported by evidence. Our framework supports this study by allowing the simulation of habitual behaviour.

Part IV contains one chapter that concludes our thesis, discusses our central choices and describes the societal and scientific relevance of my dissertation.

Relevance This thesis provides a grounded and (re)useable agent framework that supports the simulation of routines and as such is relevant for scientific work on

SPT in ABSS by emphasising domain-independence, integrating agent theory and SPT and conceptualising SPT specifically to support ABSS.

ABSS by enabling simulations where humans get stuck in behaviour that does not align with their goals (i.e., habits), make decisions based on more than their own welfare (i.e., values), reason based on a collective view on the world (i.e., sociality) and make decisions within a wider context of interconnected decisions (i.e., interconnectivity). By combining the strengths of ABSS and

socio-cognitive theory, we enable a new way to know, explore and improve the world grounded in evidence. We demonstrate this in applications of SoPrA in Part III and master theses D'Haens [65], Mercuur [167], Roes [217]).

Multi-Agent Systems, AI by enabling agents with traceable human-like decision-making. This leads to autonomous agents that use human routines, which promotes efficient decision-making and enables autonomous agents to understand and interact with human routines (see Chapter 5).

Sociology and Psychology by crystallising socio-cognitive theories and enabling exploration of these theories via simulation (see Chapter 7).

SoPrA provides theoretical groundwork for (policy) models of routine behaviour and helps to undestand and improve the role of routines in social systems. SoPrA's reuseable and evidence-driven approach promotes policies that are comparable, trustworthy, understandable and socially supported (see Section 8.2). In particular, SoPrA is relevant for

AI safety by providing a simulated training partner for autonomous agents that reproduces routine human behaviour and helps autonomous agents to successfully estimate human values (Chapter 5).

Social Routines in the ER by providing a formal model to identify social bottlenecks in the ER such as conflicting views on the values relevant for triage (see Chapter 6).

Commuting Routines by showing via simulations that large groups have a higher chance to cluster towards the same routine behaviour than small groups (see [217]) and that different habitual theories lead to different predictions regarding transport mode behaviour (see Chapter 7).

Consumption Routines by enabling research on the connection of the consumption of animal productions with shopping and cooking routines and the interplay between habits to consume animal products and the collective view on the consumption of animal products (see [167] and Section 8.2.2).

Samenvatting

Motivatie Het gedrag van groepen mensen speelt een belangrijke maar complexe rol bij een groot aantal sociale uitdagingen, zoals epidemieën, klimaatverandering en het coördineren van een ziekenhuis. Bij een epidemie is het bijvoorbeeld belangrijk te weten waar mensen heen gaan en in welke mate zij zich conformeren aan de beperking van sociale contacten. Dit is echter niet eenvoudig te achterhalen: individuen gedragen zich verschillend, beïnvloeden elkaar en passen hun gedrag in de loop van de tijd aan. Agent-gebaseerde sociale simulatie (ABSS) stelt onderzoekers in staat om deze complexe aspecten van menselijk gedrag te simuleren door individuele entiteiten (agenten genoemd) en hun interacties rechtstreeks te representeren [107]. ABSS-studies worden gebruikt om beleid te sturen, theorieën bloot te leggen, sociale fenomenen te verklaren, sociale systemen te beschrijven, of belangrijke afhankelijkheden te illustreren [107]. Kortom, ABSS-studies verschaffen en communiceren inzichten over complexe sociale systemen.

Huidige raamwerken voor agentmodellen ondersteunen onderzoekers niet in ABSS-studies naar menselijke routines (zie hoofdstuk 3). Deze raamwerken richten zich op redeneer-aspecten, sociale aspecten of psychologische mechanismen zoals angst of behoeften. Deze aspecten volstaan niet om de hoofdaspecten van menselijke routines te vatten: mensen nemen beslissingen uit gewoonte, verbinden deze gewoontes door de dag heen en gebruiken deze onderling verbonden gewoontes als een blauwdruk voor sociale interactie (zie hoofdstuk 3). Bewijsmateriaal uit de sociale wetenschappen toont de invloed van routines op menselijk beslissingsgedrag (zie hoofdstuk 3). Empirisch blijkt uit metastudies op het gebied van vervoerskeuzes [124], voedselkeuzes [215] en recycling [148] steevast dat metingen van gewoonten een gemiddelde tot sterke correlatie hebben met gedrag. Generieke theorieën voor menselijk beslissingsgedrag bevestigen deze hoofdrol van routines [8, 137, 250]. Om William James te citeren over gewoontes: "Negenennegentighonderdste of misschien wel negenhonderd-negenennegentig-duizendste van onze activiteit is puur automatisch en gewoontegedrag, van 's morgens opstaan tot 's avonds gaan liggen" [136, Hoofdstuk VIII]. Naast dit bewijs voor het belang van het gewoonte-aspect van routines zijn er vergelijkbare theoretische en empirische bewijzen voor het belang van de sociale en planning aspecten van routines (zie hoofdstuk 3). Een agent raamwerk dat menselijke routines integreert zou ABSS-onderzoekers in staat stellen nieuwe inzichten te verwerven in sociale systemen die gebaseerd zijn op deze theoretische en empirische bewijzen.

Doel Dit proefschrift levert het 'Social Practice Agent'-raamwerk, oftewel SoPrA-raamwerk, dat de volgende onderzoeksdoelstelling vervult:

Onderzoeksdoelstelling

Creëer een domein-onafhankelijk agent-raamwerk dat theorieën over sociale praktijken integreert om de simulatie van menselijke routines te ondersteunen.

Dit proefschrift gebruikt de sociale praktijktheorie (SPT) om een agentraamwerk te conceptualiseren dat onze menselijke routines integreert. SPT is een sociaal-cognitieve theorie die goed past bij het modelleren van menselijke routines, omdat deze theorie tot doel heeft ons 'dagelijks doen en zeggen' te beschrijven [227]. Door SoPrA te bouwen op een theorie uit de sociologie (en de theorie uit te breiden en te verfijnen aan de hand van verwant bewijsmateriaal), voldoet ons raamwerk aan twee criteria die het succes van ABSS-studies ondersteunen:

Gronding in bewijs zorgt ervoor dat we de relatie versterken tussen het bewijs over menselijke besluitvorming en ons model van menselijke besluitvorming: agenten. Goed gegronde agent-modellen produceren goed gegronde ABSS-studies.

Herbruikbaarheid zorgt ervoor dat de modellen die gebaseerd zijn op ons raamwerk bruikbaar zijn voor verschillende wetenschappers, vergelijkbaar zijn met andere modellen en gemakkelijk kunnen worden uitgebreid of gecombineerd met andere modellen. De herbruikbaarheid van een model is belangrijk voor het ABSS-vakgebied als collectief: zo worden we efficiënter en kunnen we modellen laten evolueren (i.e., selecteren, reproduceren, muteren) richting steeds succesvollere representaties van mensen. Om de herbruikbaarheid van onze agentmodellen te verzekeren hebben we een *domein-onafhankelijk* agentraamwerk nodig dat ABSS-onderzoekers in staat stelt om meerdere *domein-afhankelijke* agentmodellen te creëren.

Overzicht & Resultaten Deel I bevat één hoofdstuk waarin we onze motivatie, onderzoeksdoelstelling en bovengenoemde vragen verder uitwerken. Deel II van het proefschrift identificeert de aspecten van de sociale praktijktheorie die relevant zijn voor agent-gebaseerde simulatie, distilleert eisen uit de literatuur, bekijkt huidige agent-raamwerken en levert het SoPrA-raamwerk dat aan de genoemde eisen voldoet.

Hoofdstuk 2 levert een voorstudie die zich richt op twee aspecten van routines: waarden en normen. We ontdekten dat menselijke waarden een universele, morele, geoperationaliseerde, empirisch onderbouwde en een trans-contextuele basis bieden voor intentioneel gedrag en dat normen de dynamische sociale component van gedrag bieden. Waarden en normen leveren de noodzakelijke submodellen voor SoPrA en geven inzicht in de meerwaarde van de SP-concepten ten opzichte van een simpeler model met alleen waarden en normen.

Hoofdstuk 3 zoomt weer uit en beschouwt de SP als een geheel. We bekijken de literatuur uit de sociologie, psychologie en agenttheorie met betrekking tot routines. In het bijzonder:

- We vinden dat we ABSSs, de agententheorie en SPT kunnen integreren door ons te richten op drie belangrijke aspecten van gedrag: het gewoonteaspect, het sociale aspect en hoe gedrag verschillende activiteiten verbindt (dit beantwoordt **RQ1**). Deze aspecten weerspiegelen huidig bewijs uit de literatuur dat de beslissingen van mensen automatisch en contra-intentioneel zijn, gebaseerd zijn op menselijke waarden, gebaseerd zijn op een collectieve kijk op de wereld en afhankelijk zijn van hiërarchieën van activiteiten.
- We geven een lijst van vereisten voor een agent-raamwerk dat SPT integreert om simulaties van menselijke routines mogelijk te maken (dit beantwoordt **RQ2**). Deze vereisten bevorderen een duidelijke relatie tussen SoPrA en het huidige empirische en theoretische bewijs en zorgen ervoor dat SoPrA een transparant anker biedt in verder onderzoek.
- We tonen aan dat de huidige agent-raamwerken onderzoekers niet ondersteunen in het simuleren van menselijke routines (dit beantwoordt **RQ3**). In het bijzonder missen huidige agent-raamwerken (1) een ondersteuning van context-afhankelijke gewoonten, individueel leren met betrekking tot gewoonten en expliciet redeneren over gewoonten (2) een uitgebreide set van collectieve concepten, een efficiënte ordening van sociale informatie rond acties en een koppeling van individuele en collectieve concepten, en (3) expliciete relaties tussen activiteiten, tussen elke activiteit en elk ander modelconcept en de mogelijkheid hiërarchieën van activiteiten te modelleren.

Hoofdstuk 4 presenteert het SoPrA-raamwerk dat SPT integreert en voldoet aan de eisen zoals uiteengezet in hoofdstuk 3 (dit beantwoordt **RQ4**).

Deel III beschrijft drie toepassingen van SoPrA. Deze casussen laten zien hoe ons raamwerk in staat is om onderzoek in verschillende domeinen met verschillende doelen te ondersteunen door gebruik te maken van verschillende onderdelen van SoPrA (dit beantwoordt **RQ5**).

Hoofdstuk 5 presenteert een studie die zich richt op het 'waarde-harmonisatie'-probleem: een bekend probleem in AI waarbij men probeert het gedrag van een autonome agent te rijmen met menselijke waarden. De studie analyseert in welke mate een autonome agent in staat is om de waarden en normen van een gesimuleerd mens in te schatten.

Hoofdstuk 6 presenteert een studie over het identificeren van sociale knelpunten in ziekenhuizen. De studie geeft een formeel model (een OWL-ontologie) van de sociale dimensie van de spoedeisende hulp (ER) en gebruikt een formele (computer)redeneerder om die knelpunten te vinden.

Hoofdstuk 7 presenteert een studie waarin, met een simulatie, twee theorieën worden vergeleken die gewoontegedrag beschrijven. Via simulatie identificeren we een empirisch experiment dat sociale wetenschappers in staat stelt uit te zoeken welke theorie door bewijs wordt ondersteund. Ons raamwerk ondersteunt deze studie door de simulatie van gewoontegedrag mogelijk te maken.

Deel IV bevat één hoofdstuk dat ons proefschrift afsluit, ingaat op onze centrale keuzes, en de maatschappelijke en wetenschappelijke relevantie beschrijft van mijn proefschrift.

Relevantie Dit proefschrift levert een gegrond en herbruikbaar agent-raamwerk dat de simulatie van routines ondersteunt en als zodanig relevant is voor wetenschappelijk werk aan

SPT in ABSS door de nadruk te leggen op domein-onafhankelijkheid, de integratie van agenttheorie en SPT en het conceptualiseren van SPT specifiek om ABSS te ondersteunen.

ABSS door simulaties mogelijk te maken waarin mensen vastlopen in gedrag dat niet in lijn is met hun doelen (i.e., gewoonten), beslissingen nemen op basis van meer dan hun eigen welzijn (i.e., waarden), redeneren op basis van een collectieve kijk op de wereld (i.e., socialiteit) en beslissingen nemen binnen een bredere context van onderling verbonden beslissingen (i.e., interconnectiviteit). Door de sterke punten van ABSS en socio-cognitieve theorie te combineren, maken we een nieuwe manier mogelijk om de wereld te kennen, te verkennen en te verbeteren, gebaseerd op bewijsmateriaal. We tonen dit aan in applicaties van SoPrA in Part III en masterscripties [65, 167, 217]).

Multi-Agent Systemen, AI door agenten met traceerbare mensachtige besluitvorming mogelijk te maken. Dit leidt tot autonome agenten die menselijke routines gebruiken, wat efficiënte besluitvorming bevordert en autonome agenten in staat stelt menselijke routines te begrijpen en ermee te interacteren (zie hoofdstuk 5).

Sociologie en psychologie door het uitkristalliseren van sociaal-cognitieve theorieën en het mogelijk maken van exploratie van deze theorieën via simulatie (zie hoofdstuk 7).

SoPrA geeft een theoretisch gegronde basis aan (beleids)modellen van routinegedrag en helpt bij het begrijpen en verbeteren van de rol van routines in sociale systemen. De herbruikbare en bewijsgedreven aanpak van SoPrA bevordert beleid dat vergelijkbaar, betrouwbaar en begrijpelijk is en maatschappelijk wordt gedragen (zie hoofdstuk 8.2). SoPrA is in het bijzonder relevant voor

AI-veiligheid door een gesimuleerde trainingspartner voor autonome agenten te bieden die routinematig menselijk gedrag reproduceert en autonome agenten helpt om menselijke waarden met succes in te schatten (zie hoofdstuk 5).

Sociale routines bij de eerste hulp door een formeel model te bieden om sociale knelpunten op de EH te identificeren, zoals conflicterende opvattingen over de waarden die relevant zijn voor triage (zie hoofdstuk 6).

Forensroutines door via simulaties aan te tonen dat grote groepen een grotere kans hebben om naar hetzelfde routinegedrag te clusteren dan kleine groepen (zie [217]) en dat verschillende gewoontetheorieën leiden tot verschillende voorspellingen met betrekking tot vervoerskeuzegedrag (zie hoofdstuk 7).

Consumptieroutines door onderzoek mogelijk te maken naar de samenhang van de consumptie van dierlijke producten met boodschap- en kookroutines. En door onderzoek mogelijk te maken naar de wisselwerking tussen de gewoonte om dierlijke producten te consumeren en de collectieve visie op dierlijke producten te consumeren (zie [167] en sectie 8.2.2).

Acknowledgements

Catholijn, dank voor uitdragen van oprechtheid en medemenselijkheid. Dank voor het durven laten zien van je imperfecties. Dank voor mij leren schrijven. Ik zal je stokpaardjes nog jaren ongevraagd doorgeven aan anderen. Dank voor je kennis en kunde in KI. Dank voor je onophoudelijk geloof in het belang van mijn onderzoek. Dank voor het tijd maken om te wandelen en te luisteren. **Virginia**, dank voor het staan voor goed onderzoek dat bijdraagt aan een mooiere wereld. Dank dat je mij motiveerde om me hard te maken voor wat ik belangrijk vind aan mijn onderzoek. Dit was niet gemakkelijk voor iemand die als kind leerde dat de beste manier om te overleven was zich zo klein mogelijk te maken. Maar het is een noodzakelijke vaardigheid in deze wereld en het zal me goed van pas komen om te vechten voor waar ik in geloof. **Martijn**, dank voor het op weg helpen in de autoritten in het eerste jaar en dank voor het naar huis brengen het laatste jaar. Je bent een voorbeeld voor me in het plezier wat je hebt in je werk en contact met studenten. Dank voor je vertrouwen in mijn kunnen. We weten beiden dat het echte PhD-werk gedaan is in gesprekken met jou in de auto. **Robert**, dank voor je mentorschap. Dank voor alle positieve gesprekjes op de gang. Je stond altijd klaar met een goede organisatorische analyse of praktische tip. Ik ga nog jaren lang profijt hebben van je ideeën over hoe je de laag boven je managet. (Ik ga er maar even vanuit dat de laag boven me hier gestopt is met lezen.)

Charlotte, dank voor alle gesprekken in bed waar je mij geruststelde. Dank voor het meeleven, voor naast me staan, voor met mij zijn. Dank voor elke keer dat je zei dat de PhD ook jou te veel werd. Dank voor het luisteren. Dank voor al je geduld. Dank voor het mij zien los van deze promotie. Dank voor mij gronden in wat belangrijk is in het leven. Ik hou zo veel van je en ben zo blij met ons. Lieve **moeder**, dank voor je onophoudelijke steun en liefde. Dank voor het benadrukken dat het altijd het beste is je eigen mening te vormen, dat je andere mensen niet kan veroordelen omdat ze toch geen vrije wil hebben en dat een goed mens maatschappelijk betrokken is. Dank dat je altijd openstond om 'vroeger' nog een keer door te spreken. Dank voor alle liefde en wijsheid die je me hebt kunnen geven in een geweldig lastige situatie. **Vos**, dank voor je trots en steun door de jaren. Laten, we, zoals altijd, maar weer snel afspreken.

Nico, René, dank voor alle inspiratie. Dank voor het samen bedrijven van echte wetenschap: een grenzeloze nieuwsgierige en skeptische zoektocht naar de magie van het leven. **Rogier**, dank voor alle uit de hand gelopen gesprekken van modelleren tot de hero's journey, de meta, het goede leven, wiskunde en Rawls. Je bent voor mij wat een PhD-student moet zijn: je loopt je eigen pad gemotiveerd door nieuwsgierigheid en plezier met nul tolerantie voor bullshit. **Niels, Vincent, Jetze, Nina, Yfke, Susan** en al mijn andere favoriete **(Tan)CKI-vriendjes**, dank voor een thuis tijdens de studie en de afgelopen jaren. Dank aan **Wake Up** voor

alle liefde en mindfulness van het afgelopen jaar. Dank aan de 'mannenclub' (ugh), **Bernard, Ruud, Okke**, voor de verbondenheid en het vertrouwen om het leven recht in de ogen aan te kijken. **Lian**, dank voor de liefdevolle en fijne tijd samen vol met veganisme, boedhisme en wetenschap. Ik gun je de wereld en de wereld een boeddhistische psychedelische non-therapeut die hersen-autisme-onderzoek doet. Dank aan al mijn vriendjes en vriendinnetjes in Nijmegen die voor een rust-punt zorgde in de chaos van de PhD: **JH**, **Laurens**, **Lieke**, **Jan**, **Danny**, **Alex** en **Ruben**.

Dank aan al mijn geweldige collega's van de afgelopen jaren. **Ben**, wat een rust en wijsheid zit er in jou. Als nuchtere Fries moest je dat imago natuurlijk ook hoog houden. Ik land altijd in een warm bad van begrip en relativering als ik met jou praat. Thanks **Christine** for being my first real colleague. Our conversations were often the highlight of my day. I'm hoping for many more years of complaining about the scientific environment to successfully hide that it's the only place we really feel at home. **Wiebke**, thanks for all your positive energy. Let it flood the world as it flooded me. **Vladimir**, thanks for all the 'bakkies doen' and our activist extravangzas. Thanks to all the **TPM employees** who always supported me over the years, **Fabio, Mannat, Jordy, Klara, Huib, Marijn, Anneke, Mark** and **Jolien**. Thanks to all the members of **Interactive Intelligence group** for making me feel at home (and playing table tennis) when I needed it. Special thanks to **Fran** for doing an extra PhD on social cohesion work for II and **Bernd, Frank, Pietro** and **Malte** for many amazing DnD sessions and discussions. Thanks **Luciano**, our shared paper is an example of how I would like research to be. Thanks to Dirk's minions, **Amir, Arthur, Indushree, Sergei, Farzam** and **Xavier**, for being able to laugh together about all the hardship of this process.

Thanks to the generation before me in the **social simulation community** that paved the way for research that captures all the interesting, cool, weird stuff about the world. Special thanks to **Frank Dignum**, for being the main inspiration for this work. Thanks for sharing your amazing ideas so generously and freely **Frank**. Now let's hope together the world recognizes our genius soon. **Wander**, thanks for being everything a scientist should be but often isn't. **Dirk**, thanks for showing me the art of good storytelling and celebrating the importance of science. **Gary**, thanks for actually reading this whole damn thing I wrote in detail. **Gert-Jan** and **Bruce**, thanks for fulfilling my stereotype of the highly intelligent, slightly chaotic scientific mastermind. You should both get a big beard though to finish it off. Thanks **John, Patrycja, Emile, Caspar, Amineh** and all the other amazing colleagues that shared their ideas with me over the years.

Dank aan alle zorgverleners die me de afgelopen vijf jaar hebben geholpen met een gezonde geest en een gezond lichaam! In het speciaal **Henny** dank voor je steun en begrip van het afgelopen jaar. Het absurde van dit alles is hoeveel mensen een rol hebben gespeeld in de afgelopen vijf jaar. Ik moet denken aan alle gesprekjes op de gang met collega's die inmiddels alweer uit mijn leven verdwenen zijn of buurmannen van een huis waar ik niet meer woon, toneelleden van een club die ik niet meer zie. Al die gesprekjes zijn uiteindelijk waar het mij om gaat. Ubuntu!

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*. ACM, 2004.

[2] W. Abrahamse, L. Steg, C. Vlek, and T. Rothengatter. A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, 25:273–291, 2005. ISSN 02724944. doi: 10.1016/j.jenvp.2005.08.002.

[3] M. A. Adriaanse, P. M. Gollwitzer, D. T. de Ridder, J. B. de Wit, and F. M. Kroese. Breaking habits with implementation intentions: A test of underlying processes. *Personality and Social Psychology Bulletin*, 37(4):502–513, 2011. ISSN 01461672. doi: 10.1177/0146167211399102.

[4] A. Aerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. World Value Survey: Round Seven – Country-Pooled Datafile. Technical report, JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria, 2020. URL http://www.worldvaluessurvey.org/wvs.jsp.

[5] V. Albino, N. Carbonara, and I. Giannoccaro. Coordination mechanisms based on cooperation and competition within industrial districts: An agent-based computational approach. *Journal of Artificial Societies and Social Simulation*, 6(4), 2003.

[6] D. Allen. Re-reading nursing and re-writing practice: Towards an empirically based reformulation of the nursing mandate. *Nursing Inquiry*, 11(4):271–283, 2004. ISSN 13207881. doi: 10.1111/j.1440-1800.2004.00234.x.

[7] J. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[8] K. Ameriks and D. M. Clarke. *Aristotle: Nicomachean Ethics*. Cambridge University Press, 2000.

[9] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in {AI} safety. *arXiv preprint: 1606.06565*, 2016.

[10] B. A. Anderson. The attention habit: How reward learning shapes attentional selection. *Annals of the New York Academy of Sciences*, 1369(1):24–39, 2016. ISSN 17496632. doi: 10.1111/nyas.12957.

[11] J. R. Anderson. *Cognitive Psychology an its implications*. Macmillan, 2005. ISBN 9788578110796. doi: 10.1017/CBO9781107415324.004.

[12] L. R. Anderson, Y. V. Rodgers, and R. R. Rodriguez. Cultural differences in attitudes toward bargaining. *Economics Letters*, 69(1):45–54, 2000. ISSN 01651765. doi: 10.1016/S0165-1765(00)00287-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0165176500002871.

[13] G. Andrighetto and R. Conte. Emergence In the Loop: simulating the two way dynamics of norm innovation. In *Dagstuhl Seminar Proceedings*, pages 1–30, Leibniz, 2007. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. URL http://drops.dagstuhl.de/opus/volltexte/2007/907.

[14] C. J. Armitage and M. Conner. Efficacy of the Theory of Planned Behaviour: A meta-analytic review. *British Journal of Social Psychology*, 40:471–499, 2001. ISSN 0144-6665. doi: 10.1348/014466601164939.

[15] F. G. Ashby, B. O. Turner, and J. C. Horvitz. Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, 14(5):208–215, 2010. ISSN 1879307X. doi: 10.1016/j.tics.2010.02.001. URL http://dx.doi.org/10.1016/j.tics.2010.02.001.

[16] K. Atkinson and T. Bench-Capon. Value Based Reasoning and the Actions of Others. In *Proceedings of ECAI*, pages 680–688, 2016. ISBN 9781614996729. doi: 10.3233/978-1-61499-672-9-680.

[17] A. Augello, F. Dignum, M. Gentile, I. Infantino, U. Maniscalco, G. Pilato, and F. Vella. A social practice oriented signs detection for human-humanoid interaction. *Biologically Inspired Cognitive Architectures*, 25(June):8–16, 2018. ISSN 2212683X. doi: 10.1016/j.bica.2018.07.013. URL https://doi.org/10.1016/j.bica.2018.07.013.

[18] A. J. Bagnall and G. D. Smith. A multiagent model of the UK market in electricity generation. *IEEE Transactions on Evolutionary Computation*, 9(5): 522–536, 2005. ISSN 1089778X. doi: 10.1109/TEVC.2005.850264.

[19] T. Balke, T. Roberts, M. Xenitidou, and N. Gilbert. Modelling Energy-Consuming Social Practices as Agents. In Miguel, Amblard, Barceló, and Madella, editors, *Advances in Computational Social Sciene and Social Simulation*, Barcelona, 2014. Autònoma University of Barcelona.

[20] M. Batty. *The new science of cities*. MIT press, 2013.

[21] F. Benoit. Public policy models and their usefulness in public health: The stages model. *National Collaborating Centre for Healthy Public Policy*, 2013.

[22] D. Berardi, D. Calvanese, and G. D. Giacomo. Reasoning on UML Class Diagrams using Description Logic Based Systems. In *Proceedings of the KI-2001 Workshop on Applications of Description Logics (KIDLWS'01)*, pages 1–12, 2001. doi: 10.1.1.177.1920.

[23] M. Berg, J. Aarts, and J. Van der Lei. ICT in Health Care: Sociotechnical Approaches. *Methods of Information in Medicine*, 42(4):297–301, 2003. ISSN 00261270. doi: 10.1267/METH03040297.

[24] J. A. Bergstra, J. Heering, and P. Klint. Module algebra. *Journal of the ACM (JACM)*, 37(2):335–372, 1990.

[25] H. Bersini. UML for ABM. *Journal of Artificial Societies and Social Simulation*, 15(2012):1–16, 2014.

[26] W. Bilsky, M. Janik, and S. H. Schwartz. The Structural Organization of Human Values-Evidence from Three Rounds of the European Social Survey (ESS). *Journal of Cross-Cultural Psychology*, 42(5):759–776, 2011. ISSN 0022-0221. doi: 10.1177/0022022110362757.

[27] M. M. Biskjaer. An Introduction to ' Creativity Constraints ' as a Concept for Cross-Disciplinary Interchange in Creativity Research. *Innovating in Global Markets: Challenges for Sustainable Growth. Proceedings*, 2013.

[28] C. Bock, A. Fokoue, P. Haase, R. Hoekstra, I. Horrocks, A. Ruttenberg, U. Sattler, and M. Smith. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition), 12 2012. URL http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/.

[29] R. Boero and F. Squazzoni. Does empirical embeddedness matter? Methodological issues on agent-based models for analytical social science. *Journal of Artificial Societies and Social Simulation*, 8(4):1–31, 2005. ISSN 14607425.

[30] R. Boero, M. Castellani, and F. Squazzoni. Micro behavioural attitudes and macro technological adaptation in industrial districts: an agent-based prototype. *Journal of Artificial Societies and Social Simulation*, 7(2), 2004.

[31] P. Bordia, E. Hobman, E. Jones, C. Gallois, and V. J. Callan. Uncertainty during organizational change: Types, consequences, and management strategies. *Journal of Business and Psychology*, 18(4):507–532, 2004. ISSN 08893268. doi: 10.1023/B:JOBU.0000028449.99127.f7.

[32] P. Bordia, E. Jones, C. Gallois, V. J. Callan, and N. Difonzo. Management are aliens!: Rumors and stress during organizational change. *Group and Organization Management*, 31(5):601–621, 10 2006. ISSN 10596011. doi: 10.1177/1059601106286880.

[33] P. Bordia, K. Kiazad, S. L. D. Restubog, N. DiFonzo, N. Stenson, and R. L. Tang. Rumor as Revenge in the Workplace. *Group and Organization Management*, 39(4):363–388, 8 2014. ISSN 15523993. doi: 10.1177/1059601114540750.

[34] P. Bourdieu. *Outline of a theory of practice*, volume 16. Cambridge University Press, 1977. URL papers://cfc50b6a-2d9e-4feb-87e5-d6012043bd5a/Paper/p2118.

[35] M. E. Bouton. A learning theory perspective on lapse, relapse, and the maintenance of behavior change. *Health Psychology*, 19(1, Suppl):57–63, 2005. ISSN 0278-6133. doi: 10.1037/0278-6133.19.suppl1.57.

[36] M. E. Bratman. Shared cooperative activity. *The philosophical review*, 101 (2):327–341, 1992.

[37] W. F. Brewer. Bartlett's concept of the schema and its impact on theories of knowledge representation in contemporary cognitive psychology. *Bartlett, culture and cognition*, pages 69–89, 2000.

[38] J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre. The BOID architecture. In *Proceedings of the fifth international conference on Autonomous agents*, pages 9–16, 2001. ISBN 158113326X. doi: 10.1145/375735.375766.

[39] J. Bruntse. *Hospital Staff Planning with Multi-Agent Goals John Bruntse Larsen*. PhD thesis, Technical University of Denmark, 2019.

[40] H. T. Buckner. A Theory of Rumor Transmission. *Public Opinion Quarterly*, 29(1):54–70, 1965. ISSN 0033362X. doi: 10.1086/267297.

[41] K. Carley and A. Newell. The nature of the social agent, 1994. ISSN 15455874.

[42] N. Cass and J. Faulconbridge. Commuting practices: New insights into modal shift from theories of social practice. *Transport Policy*, 45:1–14, 2016. ISSN 1879310X. doi: 10.1016/j.tranpol.2015.08.002. URL http://dx.doi.org/10.1016/j.tranpol.2015.08.002.

[43] C. Castelfranchi. Modeling social action for AI agents. *IJCAI International Joint Conference on Artificial Intelligence*, 2(c):1567–1576, 1997. ISSN 10450823. doi: 10.1016/S0004-3702(98)00056-3.

[44] C. Castelfranchi. Cognitive architecture and contents for social structures and interactions. In R. Sun, editor, *Cognition and multi-agent interaction: from cognitive modeling to social simulation*, pages 349–363. Cambridge University Press, 2006. ISBN 0262201542. doi: 10.1088/0305-4470/37/26/B01. URL http://stacks.iop.org/0305-4470/37/i=26/a=B01?key=crossref.f6cb4117463d9c8d19a33f7ad17928b5.

[45] C. Castelfranchi. For a science of layered mechanisms: beyond laws, statistics, and correlations. *Theoretical and Philosophical Psychology*, 5(June): 536, 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2014.00536.

[46] L. Chen, C. D. Nugent, and H. Wang. A Knowledge-Driven Approach to Activity Recognition in Smart Homes. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):961–974, 2012. ISSN 1041-4347. doi: 10.1109/TKDE.2011.51.

[47] Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Drivers, Trends and Mitigation. Technical report, Intergovernmental Panel on Climate Change, 2015.

[48] D. J. Cooper and E. G. Dutcher. The Dynamics of Responder Behavior in Ultimatum Games: A Meta-Study. *Experimental Economics*, 14(4):519–546, 2011. ISSN 13864157. doi: 10.1007/s10683-011-9280-x.

[49] D. J. Cooper and J. H. Kagel. Other regarding preferences: a selective survey of experimental results. In J. Kagel and A. Roth, editors, *Handbook of experimental economics 2*, volume 2. Princeton university press, 2013.

[50] D. J. Cooper, N. Feltovich, A. E. Roth, and R. Zwick. Relative versus absolute speed of adjustment in strategic environments: Responder behavior in ultimatum games. *Experimental Economics*, 6(2):181–207, 2003. ISSN 13864157. doi: 10.1023/A:1025309121659.

[51] S. Cranefield, M. Winikoff, V. Dignum, and F. Dignum. No pizza for you: Value-based plan selection in BDI agents. *IJCAI International Joint Conference on Artificial Intelligence*, pages 178–184, 2017. ISSN 10450823. doi: 10.24963/ijcai.2017/26.

[52] S. Cranefield, M. Winikoff, V. Dignum, and F. Dignum. No Pizza for You: Value-based Plan Selection in BDI Agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 178–184, 2017.

[53] S. E. S. Crawford and E. Ostrom. A Grammar of Institutions. *Political Science*, 89(3):582–600, 2007. ISSN 0003-0554. doi: 10.2307/2082975. URL http://www.jstor.org/stable/2082975.

[54] R. Creath. Logical empiricism, 2011. URL https://plato.stanford.edu/entries/logical-empiricism/.

[55] D. J. Daley and D. G. Kendall. Epidemics and Rumours. *Nature*, 204(4963): 1118, 1964. ISSN 00280836. doi: 10.1038/2041118a0.

[56] E. Davidov, P. Schmidt, and S. H. Schwartz. Bringing values back in: The adequacy of the European social survey to measure values in 20 countries. *Public Opinion Quarterly*, 72(3):420–445, 2008. ISSN 0033362X. doi: 10.1093/poq/nfn035.

[57] I. de Poel and others. *Ethics, technology, and engineering: An introduction*. John Wiley & Sons, 2011.

[58] T. Deacon. *Incomplete nature: How mind emerged from matter*. WW Norton \& Company, 2011.

[59] S. Debove, N. Baumard, and J. B. Andre. Models of the evolution of fairness in the ultimatum game: A review and classification. *Evolution and Human Behavior*, 37(3):245–254, 2016. ISSN 10905138. doi: 10.1016/j.evolhumbehav.2016.01.001.

[60] F. Dechesne, G. Di Tosto, V. Dignum, and F. Dignum. No Smoking Here: Values, Norms and Culture in Multi-Agent Systems. *Artificial Intelligence and Law*, 21(1):79–107, 8 2012. ISSN 0924-8463. doi: 10.1007/s10506-012-9128-5. URL http://link.springer.com/10.1007/s10506-012-9128-5.

[61] F. Dechesne, G. Di Tosto, V. Dignum, and F. Dignum. No smoking here: values, norms and culture in multi-agent systems. *Artificial Intelligence and Law*, 21(1):79–107, 2013.

[62] G. Deffuant, F. Amblard, G. Weisbuch, and T. Faure. How can extremism prevail? A study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, 5(4), 2002.

[63] F. Del Missier, T. Mäntylä, P. Hansson, W. de Bruin, A. M. Parker, and L.-G. Nilsson. The multifold relationship between memory and decision making: An individual-differences study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(5):1344, 2013.

[64] A. K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, 2001. ISSN 16174909. doi: 10.1007/s007790170019.

[65] J. D'Haens. *Yes to agreements, yes to interventions, yes to solar panels*. PhD thesis, TU Delft, 2018. URL http://resolver.tudelft.nl/uuid:c0876137-7949-4fa8-92c6-b772965a2351.

[66] N. DiFonzo and P. Bordia. A tale of two corporations: Managing uncertainty during organizational change. *Human Resource Management*, 37(3-4):295–303, 1998. ISSN 0090-4848. doi: 10.1002/(SICI)1099-050X(199823/24)37:3/4<295::AID-HRM10>3.0.CO;2-3.

[67] N. DiFonzo, P. Bordia, and R. L. Rosnow. Reining in rumours. *Organizational Dynamics,*, 2:47–62, 1994. ISSN 0090-2616.

[68] F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999. ISSN 09248463. doi: 10.1023/A:1008315530323.

[69] F. Dignum. Interactions as Social Practices: towards a formalization. *ArXiv e-prints*, pages 1–36, 2018. URL http://arxiv.org/abs/1809.08751.

[70] F. Dignum, D. Morley, E. A. Sonenberg, and L. Cavedon. Towards socially sophisticated BDI agents. *Proceedings - 4th International Conference on MultiAgent Systems, ICMAS 2000*, pages 111–118, 2000. doi: 10.1109/ICMAS.2000.858442.

[71] F. Dignum, R. Prada, and G. Hofstede. From Autistic to Social Agents. In *Proceedings of the 2014 international conference on Autonomous Agents and Multi-Agent Systems*, pages 1161–1164, 2014. URL http://dl.acm.org/citation.cfm?id=2617431.

[72] F. Dignum, V. Dignum, R. Prada, and C. M. Jonker. A Conceptual Architecture for Social Deliberation in Multi-Agent Organizations. *Multiagent and Grid Systems*, 11(3):147–166, 2015. ISSN 18759076. doi: 10.3233/MGS-150234.

[73] V. Dignum. *A Model for Organizational Interaction*. PhD thesis, SIKS, 2004.

[74] V. Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing, 2019.

[75] V. Dignum and F. Dignum. Emergence and enforcement of social behavior. *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*, (July):2942–2948, 2009. doi: 1874/178058. URL http://www.mssanz.org.au/modsim09/H4/dignum.pdf.

[76] V. Dignum and F. Dignum. A logic of agent organizations. *Logic Journal of IGPL*, 20(September):3–7, 2012. URL http://jigpal.oxfordjournals.org/content/20/1/283.short.

[77] V. Dignum and F. Dignum. Contextualized Planning Using Social Practices. In A. Ghose, N. Oren, P. Telang, and J. Thangarajah, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 36–52. Springer International Publishing, Cham, 2015. ISBN 978-3-319-25420-3.

[78] V. Dignum, J. Vázquez-Salceda, and F. Dignum. OMNI: Introducing Social Structure , Norms and Ontologies into Agent Organizations. In *Programming MultiAgent Systems*, pages 181–198, 2005. ISBN 978-3-540-24559-9. URL http://www.springerlink.com/index/nlf2twl1uhpdn74a.pdf.

[79] J. Dubbelboer, I. Nikolic, K. Jenkins, and J. Hall. An agent-based model of flood risk and insurance. *Journal of Artificial Societies and Social Simulation*, 20(1), 2017.

[80] B. Edmonds. Bootstrapping Knowledge About Social Phenomena. *Journal of Artificial Societies and Social Simulation*, 13(2010):1–16, 2010.

[81] B. Edmonds. Different Modelling Purposes. In B. Edmonds and R. Meyer, editors, *Simulating Social Complexity*, number 9783319669472, pages 39–58. Springer, Cham, 2017. ISBN 9783319669489. doi: 10.1007/978-3-319-66948-9{_}4. URL http://link.springer.com/10.1007/978-3-319-66948-9_4.

[82] J. M. Epstein. Agent-based Computational Models and Generative Social Science. *Generative Social Science: Studies in Agent-Based Computational Modeling*, 4(5):4–46, 1999. ISSN 10762787. doi: 10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.3.CO;2-6.

[83] J. M. Epstein. *Agent_Zero: Toward neurocognitive foundations for generative social science*, volume 25. Princeton University Press, 2014.

[84] I. Erev and A. E. Roth. Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review*, 88(4):848–881, 9 1998.

[85] FAOSTAT Statistical Database. Food and Agriculture Organization of the United Nation, 2013. URL http://www.fao.org/faostat/en/.

[86] A. E. Fard, R. Mercuur, V. Dignum, C. M. Jonker, and B. v. d. Walle. Towards Agent-based Models of Rumours in Organizations: A Social Practice Theory Approach. In *Advances in Social Simulation*, pages 141–153, Cham, 2020. Springer.

[87] E. Fehr and U. Fischbacher. The nature of human altruism. *Nature*, 425 (6960):785–791, 2003. ISSN 00280836. doi: 10.1038/nature02043.

[88] P. Felli, T. Miller, C. Muise, A. R. Pearce, and L. Sonenberg. Artificial social reasoning: Computational mechanisms for reasoning about others. In *International Conference on Social Robotics*, volume 8755, pages 146–155, Cham, 2014. Springer. ISBN 9783319119724. doi: 10.1007/978-3-319-11973-1{_}15.

[89] D. Finlay and K. A. Trafimowm. The Importance of Subjective Norms for a Minority of People: between Subjects and within-Subjects Analyses. *Personality and Social Psychology Bulletin*, 22(8):820–828, 1996.

[90] M. Fishbein and I. Ajzen. *Predicting and changing behavior: The reasoned action approach*. Taylor & Francis Ltd, 2011.

[91] M. Fishbein and I. Azjen. *Predicting and Changing Behavior: The Reasoned Action Approach*. Taylor & Francis, 2011. ISBN 9781136874734.

[92] J. R. J. Fontaine, Y. H. Poortinga, L. Delbeke, and S. H. Schwartz. Structural Equivalence of the Values Domain Across Cultures: Distinguishing Sampling Fluctuations From Meaningful Variation. *Journal of Cross-Cultural Psychology*, 39(October):345–365, 2008. ISSN 0022-0221. doi: 10.1177/0022022108318112.

[93] J. Gale, K. G. Binmore, and L. Samuelson. Learning to be imperfect: The ultimatum game. *Games and Economic Behavior*, 8(1):56–90, 1995. ISSN 10902473. doi: 10.1016/S0899-8256(05)80017-X.

[94] B. Gardner. A review and analysis of the use of 'habit' in understanding, predicting and influencing health-related behaviour. *Health Psychology Review*, 9(3)(June 2014):1–19, 2014. ISSN 1743-7199. doi: 10.1080/17437199.2013.876238. URL http://www.tandfonline.com/doi/abs/10.1080/17437199.2013.876238.

[95] B. Gardner, G. J. De Bruijn, and P. Lally. A systematic review and meta-analysis of applications of the self-report habit index to nutrition and physical activity behaviours. *Annals of Behavioral Medicine*, 42(2):174–187, 2011. ISSN 08836612. doi: 10.1007/s12160-011-9282-0.

[96] M. P. Georgeff and A. L. Lansky. Reactive reasoning and planning. *Proceedings of the Sixth National on Artificial Intelligence (AAAI-87)*, pages 677–682, 1987.

[97] C. J. Gersick and J. R. Hackman. Habitual routines in task-performing groups. *Organizational Behavior and Human Decision Processes*, 47(1):65–97, 1990. ISSN 07495978. doi: 10.1016/0749-5978(90)90047-D.

[98] W. Gerstner and W. M. Kistler. Mathematical formulations of Hebbian learning. *Biological Cybernetics*, 87(5-6):404–415, 2002. ISSN 03401200. doi: 10.1007/s00422-002-0353-y.

[99] S. Ghazi, T. Khadir, and J. Dugdale. Multi-agent based simulation of environmental pollution issues: a review. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 13–21. Springer, 2014.

[100] C. Ghezzi, M. Jazayeri, and D. Mandrioli. *Fundamentals of software engineering*. Prentice Hall Englewood Cliffs, 1991.

[101] A. Ghorbani, P. Bots, V. Dignum, and G. Dijkema. MAIA: A framework for developing agent-based social simulations. *Jasss*, 16(2):1–15, 2013. ISSN 14607425. doi: 10.18564/jasss.2166.

[102] J. J. Gibson. The ecological approach to visual perception. Boston, MA, US, 1979.

[103] A. Giddens. The constitution of society: Outline of the theory of structuration. *Cognitive Therapy and Research*, 12(4):448, 1984. ISSN 0147-5916, 1573-2819. doi: 10.1007/BF01173303.

[104] R. Gifford. Environmental psychology matters. *Annual review of psychology*, 65:541–79, 2014. ISSN 1545-2085. doi: 10.1146/annurev-psych-010213-115048. URL http://www.ncbi.nlm.nih.gov/pubmed/24050189.

[105] G. N. Gilbert. *Agent-Based Models*. SAGE Publications, no. 153 edition, 2008. ISBN 1412949645. URL https://books.google.com/books?hl=nl&lr=&id=Z3cp0ZBK9UsC&pgis=1.

[106] N. Gilbert. When Does Social Simulation Need Cognitive Models? In *Cognition and Multi-Agent Interaction*, pages 428–432. Cambridge University Press, Cambridge, 12 2005. ISBN 9780511610721. doi: 10.1017/CBO9780511610721.020. URL https://www.cambridge.org/core/product/identifier/CBO9780511610721A029/type/book_part.

[107] N. Gilbert. *Agent-Based Models*. Sage Publications, Incorperated, 2019.

[108] N. Gilbert, P. Ahrweiler, P. Barbrook-Johnson, K. P. Narasimhan, and H. Wilkinson. Computational modelling of public policy: Reflections on practice. *Journal of Artificial Societies and Social Simulation*, 21(1), 2018. ISSN 14607425. doi: 10.18564/jasss.3669.

[109] H. C. J. Godfray, P. Aveyard, T. Garnett, J. W. Hall, T. J. Key, J. Lorimer, R. T. Pierrehumbert, P. Scarborough, M. Springmann, and S. A. Jebb. Meat consumption, health, and the environment. *Science*, 361(6399), 2018.

[110] L. R. Goldberg. The Structure of Phenotypic Personality Traits. *American Psychologist*, 48(1):26–34, 1993. ISSN 0003066X. doi: 10.1037/0003-066X.48.1.26.

[111] D. Goleman. *Social Intelligence*. Random House, 2011.

[112] T. R. Green and M. Petre. Usability analysis of visual programming environments: A 'cognitive dimensions' framework. *Journal of Visual Languages and Computing*, 7(2):131–174, 1996. ISSN 1045926X. doi: 10.1006/jvlc.1996.0009.

[113] E. Guerci, M. A. Rastegar, and S. Cincotti. Agent-based modeling and simulation of competitive wholesale electricity markets. In *Handbook of power systems II*, pages 241–286. Springer, 2010.

[114] W. Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3(4): 367–388, 1982. ISSN 01672681. doi: 10.1016/0167-2681(82)90011-7.

[115] W. Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, 3(4): 367–388, 1982.

[116] D. Hadfield-Menell, S. Milli, P. Abbeel, S. J. Russell, and A. Dragan. Inverse reward design. In *Proceeding of the 31st Conference on Neural Information Processing Systems (NIPS)*, pages 6765–6774, 2017.

[117] Y. Hamaguchi. Does Observation of Others Affect People's Cooperative Behavior? An Experimental Study on Threshold Public Goods Games. *Osaka Economic Papers*, 54(2):46–81, 2004. ISSN 04734548. doi: 10.1007/s001820050102.

[118] S. Harris and A. Seaborne. SPARQL 1.1 Query Language, 2013. URL https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[119] B. Hart. The psychology of rumor. *Psychiatry: Interpersonal and Biological Processes*, 28:1–26, 1916. ISSN 0035-9157.

[120] R. Hegselmann. Thomas C. Schelling and James M. Sakoda: The intellectual, technical, and social history of a model. *Jasss*, 20(3), 2017. ISSN 14607425. doi: 10.18564/jasss.3511.

[121] D. Helbing and S. Balietti. Agent-Based Modelling. In *Social Self-Organization*, pages 101–114. Springer-Verlag, 2012. ISBN 978-3-642-24003-4. doi: 10.1007/978-3-642-24004-1. URL http://link.springer.com/10.1007/978-3-642-24004-1.

[122] B. J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. Mcelreath. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review*, 91(2), 2001.

[123] K. V. Hindriks. Programming Rational Agents in GOAL. In *Multi-Agent Programming*, chapter 4. Springer, Boston, 2009. ISBN 9780387892993. doi: 10.1007/978-0-387-89299-3.

[124] C. Hoffmann, C. Abraham, M. P. White, S. Ball, and S. M. Skippon. What cognitive mechanisms predict travel mode choice? A systematic review with meta-analysis. *Transport Reviews*, 0(0):1–22, 2017. ISSN 0144-1647. doi: 10.1080/01441647.2017.1285819. URL https://www.tandfonline.com/doi/full/10.1080/01441647.2017.1285819.

[125] G. J. Hofstede. GRASP agents: social first, intelligent later. *AI and Society*, 0 (0):1–9, 2017. ISSN 14355655. doi: 10.1007/s00146-017-0783-7. URL http://dx.doi.org/10.1007/s00146-017-0783-7.

[126] G. J. Hofstede, C. Jonker, and T. Verwaart. An agent model for the influence of culture on bargaining. In *HUCOM 2008 - 1st International Working Conference on Human Factors and Computational Models in Negotiation*, pages 39–46, 2008. ISBN 9789081381116. doi: 10.1145/1609170.1609175.

[127] G. Holtz. Generating Social Practices. *Journal of Artificial Societies and Social Simulation*, 17(1):17, 2014. ISSN 1460-7425.

[128] I. Horrocks. Description Logic : A Formal Foundation for Ontology Languages and Tools, 2007.

[129] I. Horrocks, P. F. Patel-schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean. SWRL : A Semantic Web Rule Language Combining OWL and RuleML. *W3C Member submission*, (May 2004):1–20, 2004.

[130] L. Hunsberger. Algorithms for a Temporal Decoupling Problem in Multi-Agent Planning. *Aaai-02*, pages 468–475, 2002.

[131] G. Iori. Expectation formation in Agent Based Models of financial, credit and good markets, 2016. URL http://www.ssc2016.cnr.it/file_download/5/talk_new_Roma_2016_final.pdf.

[132] G. Irving and A. Askell. AI Safety Needs Social Scientists. *Distill*, 4(2):e14, 2019.

[133] W. Jager. *Modelling consumer behaviour*. The Netherlands: Universal Press, 2000. ISBN 9076269173.

[134] W. Jager. Enhancing the Realism of Simulation (EROS): On Implementing and Developing Psychological Theory in Social Simulation. *Journal of Artificial Societies and Social Simulation*, 20(3):14, 2017. ISSN 14607425. doi: DOI: 10.18564/jasss.3522.

[135] W. Jager and M. Janssen. An updated conceptual framework for integrated modeling of human decision making: The Consumat II. *Eccs 2012*, page 10, 2012.

[136] W. James. *Talks to teachers on psychology and to students on some of life's ideals*, volume 12. Harvard University Press, 1983. URL https://www.uky.edu/~eushe2/Pajares/tt8.html.

[137] W. James. *The principles of psychology*, volume 1. Cosimo, Inc., 2007. URL http://psychclassics.yorku.ca/James/Principles/prin4.htm.

[138] W. Jefferson. Complexity A more recent conceptualization of how to look at nature and our interaction with it, 2017.

[139] John Stuart Mill. On the definition of political economy, and on the method of investigation proper to it. In *Essays on some unsettled questions of Political Economy*, page 326. Routledge, 1844.

[140] C. M. Jonker and J. Treur. Formal Analysis of Models for the Dynamics of Trust based on Experiences. In *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pages 221–231, Berlin, 1999. Springer. ISBN 978-3-540-66281-5. doi: 10.1007/3-540-48437-X. URL http://link.springer.com/10.1007/3-540-48437-X.

[141] C. M. Jonker, M. B. Van Riemsdijk, and B. Vermeulen. Shared mental models. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, number January 2010, pages 132–151. Springer, 2010. ISBN 9783642212673, 3642212670. doi: 10.1007/978-3-642-21268-0{_}8.

[142] G. T. Jun, F. Carvalho, and N. Sinclair. Ethical Issues in Designing Interventions for Behavioural Change. *DRS2018: Catalyst*, 1, 2018. doi: 10.21606/drs.2018.498.

[143] J. Kagel and A. Roth. *Handbook of Experimental Economics*, volume 2. Princeton university press, 2013.

[144] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011. ISBN 9781429969352. URL http://books.google.com/books?id=ZuKTvERuPG8C.

[145] G. Kaminka. Curing robot autism: A challenge. *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, pages 801–804, 2013.

[146] A. J. Kimmel. *Rumors and rumor control : a manager's guide to understanding and combatting rumors*. Lawrence Erlbaum, Mahwah, NJ :, 2004. ISBN 9781410609502.

[147] G. Klein, K. Associates, and D. Ara. Naturalistic Decision Making. *Human Factors*, 50(3):456–460, 2008. ISSN 0018-7208. doi: 10.1518/001872008X288385.

[148] C. A. Klöckner and I. O. Oppedal. General Versus Domain Specific Recycling Behaviour. *Ressources, Conservation & Recycling*, 55(7491):463–471, 2011.

[149] R. H. Knapp. A Pyschology of Rumour. *Public Opinion Quarterly*, 8(1):22–37, 1944. ISSN 0033-362X, 1537-5331. doi: 10.1086/265665.

[150] J. Knobe and B. Malle. The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, 33(33):101–121, 1997.

[151] T. A. Knopf. Beating the Rumors: An Evaluation of Rumor Control Centers. *Policy Analysis*, 1(4):599–612, 1975. ISSN 0098-2067.

[152] M. J. Kollingbaum and T. J. Norman. Norm adoption and consistency in the NoA agent architecture, 2004. ISSN 03029743.

[153] S. Kumar, M. J. Huber, P. R. Cohen, and D. R. McGee. Toward a formalism for conversation protocols using joint intention theory. *Computational Intelligence*, 18(2):174–228, 2002. ISSN 0824-7935. doi: 10.1111/1467-8640.00187.

[154] T. Kurz and B. Gardner. Habitual behaviors or patterns of practice? Explaining and changing repetitive climate relevant actions. *Wiley Interdisciplinary Reviews: Climate Change*, 6(1):113–128, 11 2015. ISSN 17577780. doi: 10.1002/wcc.327. URL http://doi.wiley.com/10.1002/wcc.327http://onlinelibrary.wiley.com/doi/10.1002/wcc.327/full.

[155] P. Lally, C. H. M. Van Jaarsveld, H. W. W. Potts, and J. Wardle. How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology*, 40(6):998–1009, 2010. ISSN 00462772. doi: 10.1002/ejsp.674.

[156] P. Lally, J. Wardle, and B. Gardner. Experiences of habit formation: a qualitative study. *Psychology, health & medicine*, 16(4):484–489, 2011. ISSN 1354-8506. doi: 10.1080/13548506.2011.555774.

[157] B. Latour. On actor-network theory: A few clarifications, 1996. URL https://www.jstor.org/stable/40878163.

[158] S. Levine, Z. Popovic, and V. Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *Proceeding of the 31st Conference on Neural Information Processing Systems (NIPS)*, pages 19–27, 2011.

[159] E. A. Locke and G. P. Latham. *A theory of goal setting & task performance.* Prentice-Hall, Inc, 1990.

[160] A. Machado. Learning the Temporal Dynamics of Behavior. *Psychological review*, 104(2):241–265, 1997.

[161] B. F. Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press, 2006.

[162] R. S. Mans, W. M. P. V. D. Aalst, and R. J. B. Vanwersch. *Process Mining in Healthcare Evaluating and Exploiting Operational Healthcare Processes*. Springer, 2015. ISBN 9783319160702. doi: 10.1007/978-3-319-16071-9.

[163] C. F. Manski. The structure of random utility models. *Theory and decision*, 8 (3):229, 1977.

[164] J. Marschak. Binary choice constraints and random utility indicators. Technical report, YALE UNIV NEW HAVEN CT COWLES FOUNDATION FOR RESEARCH IN ECONOMICS, 1959.

[165] A. H. Maslow. A theory of human motivation. *Psychological Review*, 1943. doi: 10.4324/9781912282517.

[166] D. C. McClelland. *Human motivation*. CUP Archive, 1987.

[167] R. Mercuur. *Interventions on Contextualized Decision Making : an Agent-Based Simulation Study*. PhD thesis, Utrecht University, 2015. URL https://dspace.library.uu.nl/handle/1874/323482.

[168] R. Mercuur. Realistic Agents with Social Practices. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1752–1754. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[169] R. Mercuur, F. Dignum, and Y. Kashima. Changing Habits Using Contextualized Decision Making. In *Advances in Intelligent Systems and Computing*, volume 528, 2017. ISBN 9783319472522. doi: 10.1007/978-3-319-47253-9{_}23.

[170] R. Mercuur, V. Dignum, and C. Jonker. Modeling Social Preferences with Values. In *The Proceedings of the 3nd International Workshop on Smart Simulation and Modelling for Complex Systems @ IJCAI*, pages 17–28, 2017.

[171] R. Mercuur, V. Dignum, and C. Jonker. Using Values and Norms to Model Realistic Social Agents. In *29th Benelux Conference on Artificial Intelligence*, page 357, 2017.

[172] R. Mercuur, V. Dignum, and C. M. Jonker. Modelling Agents Endowed with Social Practices: Static Aspects. *ArXiv e-prints*, pages 1–22, 11 2018. URL http://arxiv.org/abs/1811.10981.

[173] R. Mercuur, J. B. Larsen, and V. Dignum. Modelling the social practices of an emergency room to ensure staff and patient wellbeing. *Belgian/Netherlands Artificial Intelligence Conference*, pages 133–147, 2018. ISSN 15687805.

[174] R. Mercuur, V. Dignum, C. Jonker, and others. The value of values and norms in social simulation. *Journal of Artificial Societies and Social Simulation*, 22 (1):1–9, 2019.

[175] R. Mercuur, V. Dignum, and C. M. Jonker. Discerning Between Two Theories on Habits with Agent-based Simulation. In N. Eggert, editor, *Proceedings of the Social Simulation Conference 2019*, page (in press), Mainz, 2019. Springer LNCS.

[176] R. Mercuur, V. Dignum, and C. Jonker. Integrating Social Practice Theory in Agent-Based Models: A Review of Theories and Agents. *IEEE Transactions on Computational Social Systems*, 2020.

[177] R. Mercuur, V. Dignum, and C. Jonker. Modelling human routines: Conceptualising social practice theory for agent-based simulation. *arXiv*, page (under review), 2020. ISSN 23318422.

[178] T. Metzinger and V. Gallese. The emergence of a shared action ontology: Building blocks for a theory. *Consciousness and Cognition*, 12(4):549–571, 2003. ISSN 10538100. doi: 10.1016/S1053-8100(03)00072-2.

[179] B. Meyer. *Object-oriented software construction*, volume 2. Prentice hall New York, 1988.

[180] J.-J. C. Meyer and W. Van Der Hoek. *Epistemic logic for AI and computer science*, volume 41. Cambridge University Press, 2004.

[181] G. Michelson and S. Mouly. Rumour and gossip in organisations: a conceptual study. *Management Decision*, 38(5):339–346, 6 2000. ISSN 0025-1747. doi: 10.1108/00251740010340508.

[182] A. Miles. The (Re)genesis of Values: Examining the Importance of Values for Action. *American Sociological Review*, 80(4):680–704, 2015. ISSN 0003-1224. doi: 10.1177/0003122415591800. URL http://asr.sagepub.com/content/80/4/680.abstract.

[183] T. Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *ArXiv e-prints*, 2017. doi: arXiv:1706.07269v1. URL http://arxiv.org/abs/1706.07269.

[184] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2018.

[185] S. Mindermann and S. Armstrong. Occam's razor is insufficient to infer the preferences of irrational agents. In *Conference on Neural Information Processing Systems (NIPS)*, pages 5598–5609, 2018.

[186] A. Moors and J. De Houwer. Automaticity: a theoretical and conceptual analysis. *Psychological bulletin*, 132(2):297–326, 2006. ISSN 0033-2909. doi: 10.1037/0033-2909.132.2.297.

[187] S. Moss and B. Edmonds. Towards good social science. *JASSS*, 8(4):1–16, 2005. ISSN 14607425.

[188] S. Moss and B. Edmonds. Sociology and Simulation: Statistical and Qualitative Cross-Validation. *American journal of sociology*, 110(4):1095–1131, 2005.

[189] M. A. Musen. The protégé project: a look back and a look forward. *{AI} Matters*, 1(4):4–12, 2015. doi: 10.1145/2757001.2757003. URL https://doi.org/10.1145/2757001.2757003.

[190] R. B. Myerson. *Game theory*. Harvard university press, 2013.

[191] K. Narasimhan, T. Roberts, M. Xenitidou, and N. Gilbert. Using ABM to Clarify and Refine Social Practice Theory. In W. Jager, editor, *Advances in Social Simulation 2015*, volume 528. Springer International Publishing AG 2017, 2017. ISBN 978-3-319-47252-2. doi: 10.1007/978-3-319-47253-9. URL http://link.springer.com/10.1007/978-3-319-47253-9.

[192] D. T. Neal, W. Wood, J. S. Labrecque, and P. Lally. How do habits guide behavior? Perceived and actual triggers of habits in daily life. *Journal of Experimental Social Psychology*, 48(2):492–498, 2012. ISSN 00221031. doi: 10.1016/j.jesp.2011.10.011. URL http://dx.doi.org/10.1016/j.jesp.2011.10.011.

[193] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 374(1):457–470, 2007. ISSN 03784371. doi: 10.1016/j.physa.2006.07.017.

[194] T. D. Nielsen and F. V. Jensen. Learning a decision maker's utility function from (possibly) inconsistent behavior. *Artificial Intelligence*, 160(1-2):53–78, 2004.

[195] B. Nooteboom. Agent-based simulation of trust. In *Handbook of research methods on trust*. Edward Elgar Publishing, 2015.

[196] E. Norling. Don't Lose Sight of the Forest: Why the Big Picture of Social Intelligence is Essential. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, number May, pages 984–987, 2016.

[197] D. A. Norman and T. Shallice. Attention to action. In *Consciousness and self-regulation*, pages 1–18. Springer, Boston, 1986.

[198] M. J. North, N. T. Collier, J. Ozik, E. R. Tatara, C. M. Macal, M. Bragen, and P. Sydelko. Complex adaptive systems modeling with Repast Simphony. *Complex Adaptive Systems Modeling*, 1(1):3, 2013. ISSN 2194-3206. doi: 10.1186/2194-3206-1-3. URL http://casmodeling.springeropen.com/articles/10.1186/2194-3206-1-3.

[199] E. Nouri, K. Georgila, and D. Traum. Culture-specific models of negotiation for virtual characters: multi-attribute decision-making based on culture-specific values. *AI & society*, 32(1):51–63, 2017.

[200] O. Oh, M. Agrawal, and H. R. Rao. Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises. *MIS Quarterly*, 37(2):407–426, 2013. ISSN 02767783. doi: 10.25300/MISQ/2013/37.2.05.

[201] G. Okeyo, L. Chen, and H. Wang. Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes. *Future Generation Computer Systems*, 39:29–43, 2014. ISSN 0167739X. doi: 10.1016/j.future.2014.02.014.

[202] B. Onarheim and S. Wiltschnig. Opening and constraining: constraints and their role in creative processes. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, pages 83–89, 2010.

[203] H. Oosterbeek, R. Sloof, and G. van de Kuilen. Cultural Differences In Ultimatum Game Experiments: Evidence From A Meta-Analysis. *SSRN Electronic Journal*, 188:171–188, 2001. ISSN 1556-5068. doi: 10.2139/ssrn.286428. URL http://www.ssrn.com/abstract=286428.

[204] D. C. Parker, S. M. Manson, M. A. Janssen, M. J. Hoffmann, and P. Deadman. Multi-agent systems for the simulation of land-use and land-cover change: A review. *Annals of the Association of American Geographers*, 93(2):314–337, 2003. ISSN 00045608. doi: 10.1111/1467-8306.9302004.

[205] J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, 2019.

[206] D. Pereira, E. Oliveira, and N. Moreira. Formal modelling of emotions in BDI agents. In Springer, editor, *International Workshop on Computational Logic in Multi-Agent Systems*, volume 5056 LNAI, pages 62–81, Berlin, 2007. ISBN 3540888322. doi: 10.1007/978-3-540-88833-8-4.

[207] K. Pigmans. *Value Deliberation: Towards mutual understanding of stakeholder perspectives in policymaking*. PhD thesis, Delft University of Technology, 2020.

[208] L. R. Planken, M. M. de Weerdt, and C. Witteveen. Optimal temporal decoupling in multiagent systems. In *Proceedings of the 9th International Conference on Autonomous Agents and Multi-Agent Systems*, number January, pages 789–796, 2010. doi: 10.1145/1838206.1838311.

[209] I. v. d. Poel and L. Royakkers. *Ethics, Technology, and Engineering: An Introduction*. John Wiley & Sons, 2011. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-EHEP002302.html.

[210] J. Prasad. The Psychology of Rumour: A Study Relating to the Great Indian Earthquake of 1934. *British Journal of Psychology. General Section*, 26(1):1–15, 1935. ISSN 20448295. doi: 10.1111/j.2044-8295.1935.tb00770.x.

[211] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *The Behabioral and Brain Sciences*, 4:515–526, 1978. ISSN 00491225. doi: 10.1016/j.celrep.2011.1011.1001.7.

[212] D. V. Pynadath and S. C. Marsella. PsychSim: Modeling theory of mind with decision-theoretic agents. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1181–1186, 2005. ISSN 10450823.

[213] A. Rao and M. Georgeff. BDI Agents: From Theory to Practice. *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, 1995.

[214] A. Reckwitz. Toward a theory of social practices: A development in culturalist theorizing. *European Journal of Social Theory*, 5(2):243–263, 2002. ISSN 1368-4310. doi: 10.1177/13684310222225432.

[215] J. Riet, S. J. Sijtsema, H. Dagevos, and G. J. de Bruijn. The importance of habits in eating behaviour. An overview and recommendations for future research. *Appetite*, 57(3):585–596, 2011. ISSN 01956663. doi: 10.1016/j.appet.2011.07.010.

[216] P. Ringler, D. Keles, and W. Fichtner. Agent-based modelling and simulation of smart electricity grids and markets - A literature review. *Renewable and Sustainable Energy Reviews*, 57:205–215, 2016. ISSN 18790690. doi: 10.1016/j.rser.2015.12.169. URL http://dx.doi.org/10.1016/j.rser.2015.12.169.

[217] N. Roes. *Belief Dynamis Driven By Social Influence: A Social Practice-Based Model Of Belief Adaption Driven By Intragroup Interaction With Application to Transport Mode Choices*. PhD thesis, Delft University of Technology, 2019. URL http://resolver.tudelft.nl/uuid:bfdc3a61-62a4-43b4-95f3-abee2a06e945.

[218] N. J. Roese. Counterfactual thinking. *Psychological bulletin*, 121(1):133, 1997.

[219] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro. Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, 61: 224–236, 2016. ISSN 15320464. doi: 10.1016/j.jbi.2016.04.007. URL http://dx.doi.org/10.1016/j.jbi.2016.04.007.

[220] K. Romdenh-Romluc. Habit and Attention. In *The Phenomenology of Embodied Subjectivity. Contributions to Phenomenology*, pages 159–167. Springer, Cham, 2013. ISBN 1074301491620. doi: 10.1007/s10743-014-9162-0. URL http://link.springer.com/10.1007/s10743-014-9162-0.

[221] A. Roth, V. Prasnikar, M. Okuno-fujiwara, and S. Zamir. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review*, pages 1068–1095, 1991.

[222] A. E. Roth and I. Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8(1):164–212, 1995. ISSN 10902473. doi: 10.1016/S0899-8256(05)80020-X.

[223] J. Rumbaugh, I. Jacobson, and G. Booch. *The unified modeling language reference manual*. Addison-Wesley, 2005. ISBN 0321245628. URL https://dl.acm.org/citation.cfm?id=993859.

[224] S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015. ISSN 07384602. doi: 10.1609/aimag.v36i4.2577.

[225] S. Sardina, L. De Silva, and L. Padgham. Hierarchical planning in BDI agent programming languages: A formal approach. *Proceedings of the International Conference on Autonomous Agents*, 2006:1001–1008, 2006. doi: 10.1145/1160633.1160813.

[226] T. Schatzki. On Practice Theory, or What's Practices Got to Do (Got to Do) with It? In *Education in an Era of Schooling*, pages 151–165. Springer Singapore, 2018. ISBN 978-981-13-2052-1. doi: 10.1007/978-981-13-2053-8. URL http://link.springer.com/10.1007/978-981-13-2053-8.

[227] T. R. Schatzki. *Social Practices. A Wittgensteinian approach to human activity and the social*. Cambridge University Press, 1996.

[228] T. C. Schelling. Dynamics Model of Segregation. *Journal of Mathematical Sociology*, 1(May 1969):143–186, 1971.

[229] S. H. Schwartz. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2:1–20, 2012. ISSN 2307-0919. doi: http://dx.doi.org/http://dx.doi.org/10.9707/2307-0919.1116.

[230] S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, O. Dirilen-Gumus, and M. Konty. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688, 2012. ISSN 1939-1315. doi: 10.1037/a0029393.

[231] J. Searle. Collective intentions and actions. *Intentions in communication*, page 401, 1990.

[232] J. Searle. The Construction of Social Reality, 1995.

[233] M. Seigerman. Manipulating Meat-Eating Justifications: A Novel Approach to Influencing Meat Consumption, 2014.

[234] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007. ISSN 00043702. doi: 10.1016/j.artint.2006.02.006.

[235] E. Shove, M. Pantzar, and M. Watson. *The Dynamics of Social Practice: Everyday Life and How it Changes*. SAGE Publications, London, 2012. ISBN 9780857020437. doi: 10.4135/9781446250655.n1.

[236] E. Shove, M. Watson, and N. Spurling. Conceptualizing connections: Energy demand, infrastructures and social practices. *European Journal of Social Theory*, 18(3):274–287, 2015. ISSN 1368-4310. doi: 10.1177/1368431015579964. URL http://journals.sagepub.com/doi/10.1177/1368431015579964.

[237] L. C. Siebert, R. Mercuur, V. Dignum, J. van den Hoven, and C. Jonker. Improving Confidence in the Estimation of Values and Norms. In A. Aler Tubella, S. Cranefield, C. Frantz, F. Meneguzzi, and W. Vasconcelos, editors, *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIII*, pages 98–113, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72376-7.

[238] B. Y. R. Slonim and A. E. R. I. Learning in High Stakes Ultimatum Games : An Experiment in the Slovak Republic. *The Econometric Society Stable*, 66 (3):569–596, 1998.

[239] N. Soares and B. Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, pages 103–125. Springer, 2017.

[240] A. Solaki, F. Berto, and S. Smets. The Logic of Fast and Slow Thinking. *Erkenntnis*, 2019. ISSN 15728420. doi: 10.1007/s10670-019-00128-z. URL https://doi.org/10.1007/s10670-019-00128-z.

[241] C. M. Sørup. *Development of a Generic Performance Measurement Model in an Emergency Department*. PhD thesis, Technical University of Denmark, 2015.

[242] M. Springmann, H. C. J. Godfray, M. Rayner, and P. Scarborough. Analysis and valuation of the health and climate change cobenefits of dietary change. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15):4146–4151, 2016. ISSN 10916490. doi: 10.1073/pnas.1523119113.

[243] F. Squazzoni, W. Jager, and B. Edmonds. Social Simulation in the Social Sciences: A Brief Overview. *Social Science Computer Review*, 32(3):279–294, 12 2013. ISSN 0894-4393. doi: 10.1177/0894439313512975. URL http://ssc.sagepub.com/cgi/doi/10.1177/0894439313512975.

[244] F. Squazzoni, J. G. Polhill, B. Edmonds, P. Ahrweiler, P. Antosz, G. Scholz, E. Chappin, M. Borit, H. Verhagen, F. Giardini, and N. Gilbert. Computational models that matter during a global pandemic outbreak: A call to action. *Jasss*, 23(2), 2020. ISSN 14607425. doi: 10.18564/jasss.4298.

[245] H. Storf, M. Becker, and M. Riedl. Rule-based activity recognition framework: Challenges, technique and learning. *Proceedings of the 3d International ICST Conference on Pervasive Computing Technologies for Healthcare*, 2009. doi: 10.4108/ICST.PERVASIVEHEALTH2009.6108. URL http://eudl.eu/doi/10.4108/ICST.PERVASIVEHEALTH2009.6108.

[246] T. J. Strader, F.-r. Lin, M. J. Shaw, A. Societies, and S. Simulation. Simulation of Order Fulfillment in Divergent Assembly Supply Chains. *Simulation of Order Fulfillment in Divergent Assembly Supply Chains*, 1(2):1–5, 1998. ISSN 1460-7425.

[247] C. Taylor. Interpretation and the Sciences of Man. In *Explorations in Phenomenology*, volume 25, pages 47–101. Philosophy Education Society Inc., 1973. ISBN 0034-6632. doi: 10.1007/978-94-010-1999-6{_}3.

[248] R. H. Thaler and C. R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[249] E. Thorndike. Intelligence and its uses. *Harper's Magazine*, 140:227–235, 1920.

[250] H. C. Triandis. Values, attitudes, and interpersonal behavior. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, 27:195–259, 1 1980. ISSN 0146-7875. URL http://www.ncbi.nlm.nih.gov/pubmed/7242748.

[251] N. Turenne. The rumour spectrum. *PLoS ONE*, 13(1):e0189080, 1 2018. ISSN 19326203. doi: 10.1371/journal.pone.0189080.

[252] C. Urban. PECS: A Reference Model for the Simulation of Multi-Agent Systems. *Tools and Techniques for Social Science Simulation*, pages 83–114, 2000. doi: 10.1007/978-3-642-51744-0{_}6.

[253] T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse. Accurate activity recognition in a home setting. *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp '08*, page 1, 2008. doi: 10.1145/1409635.1409637. URL http://portal.acm.org/citation.cfm?doid=1409635.1409637.

[254] M. B. Van Riemsdijk, M. Dastani, J. J. C. Meyer, and F. S. De Boer. Goal-oriented modularity in agent programming. *Proceedings of the International Conference on Autonomous Agents*, 2006:1271–1278, 2006. doi: 10.1145/1160633.1160864.

[255] B. Verplanken. Habits and implementation intentions. In *The ABC of behavioural change.*, pages 99–109. Elsevier, Oxford, UK, 2005. URL http://opus.bath.ac.uk/9438/.

[256] B. Verplanken and S. Orbell. Reflections on Past Behavior: A Self-Report Index of Habit Strength. *Journal of Applied Social Psychology*, 33(6):1313–1330, 2003. ISSN 00219029. doi: 10.1111/j.1559-1816.2003.tb01951.x.

[257] B. Verplanken, H. Aarts, A. van Knippenberg, and C. van Knippenberg. Attitude Versus General Habit: Antecedents of Travel Mode Choice. *Journal of Applied Social Psychology*, 24(4):285–300, 1994. ISSN 15591816. doi: 10.1111/j.1559-1816.1994.tb00583.x.

[258] A. Visser. Theorieën vanbinnen en vanbuiten, 2015.

[259] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 3 2018. ISSN 10959203. doi: 10.1126/science.aap9559.

[260] A. VUGTS, M. VAN DEN HOVEN, E. DE VET, and M. VERWEIJ. How autonomy is understood in discussions on the ethics of nudging. *Behavioural Public Policy*, 4(1):108–123, 2020. ISSN 2398-063X. doi: 10.1017/bpp.2018.5.

[261] W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview (Second Edition), 12 2012. URL http://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

[262] C. Wang, Z. X. Tan, Y. Ye, L. Wang, K. H. Cheong, and N.-g. Xie. A rumor spreading model based on information entropy. *Scientific Reports*, 7(1):9615, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-09171-8.

[263] F. Y. Wang, P. Ye, and J. Li. Social Intelligence: The Way We Interact, the Way We Go. *IEEE Transactions on Computational Social Systems*, 6(6):1139–1146, 2019. ISSN 2329924X. doi: 10.1109/TCSS.2019.2954920.

[264] J. J. Waring and S. Bishop. Lean healthcare: Rhetoric, ritual and resistance. *Social Science and Medicine*, 71(7):1332–1340, 2010. ISSN 02779536. doi: 10.1016/j.socscimed.2010.06.028. URL http://dx.doi.org/10.1016/j.socscimed.2010.06.028.

[265] T. v. d. Weide. Arguing to motivate decisions. *SIKS Dissertation Series*, 2011. URL http://www.narcis.nl/publication/Language/EN/id/170/RecordID/oai:dspace.library.uu.nl:1874/210788.

[266] L. Wetzel. Types and Tokens. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.

[267] W. Wood and D. T. Neal. A new look at habits and the habit-goal interface. *Psychological review*, 114(4):843–863, 2007. ISSN 0033-295X. doi: 10.1037/0033-295X.114.4.843.

[268] W. Wood, J. M. Quinn, and D. A. Kashy. Habits in everyday life: Thought, emotion, and action. *Journal of Personality and Social Psychology*, 83(6):1281–1297, 2002. ISSN 00223514. doi: 10.1037/0022-3514.83.6.1281.

[269] G. Woodruff. Premack and Woodruff : Chimpanzee theory of mind. *Behavioral and Brain Sciences*, 1(1978):515–526, 1978.

[270] M. Wooldridge and R. Jennings. Intelligent Agents: Theory and Practice. *The knowledge engineering review*, 10(2):115–152, 1995.

[271] G. H. Wright. Deontic logic. *Mind*, 60(237):1–15, 1951. doi: 10.2307/2251395.

[272] D. H. Zanette. Dynamics of rumor propagation on small-world networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 65(4):9, 3 2002. ISSN 1063651X. doi: 10.1103/PhysRevE.65.041908.