DISSERTATION

Simulation of Foundational Human Information-Processing in Social Context

Using the Concept of a State Indicator to Decide Socially Influenced Cooperative and Environmentally Friendly Behavior with the SiMA-C Mental Architecture

Submitted at the Faculty of Electrical Engineering and Information Technology, TU Wien, in partial fulfillment of the requirements for the degree of Doktor der technischen Wissenschaften (equals Ph.D.)

under supervision of

em.o.Univ.-Prof. Dipl.-Ing. Dr.techn. Dietmar Dietrich
Institute of Computer Technology, TU Wien, Vienna, Austria

and

Prof. Dr. Cristiano Castelfranchi
Institute of Cognitive Sciences and Technologies, Rome, Italy

by

Dipl.-Ing. Samer Schaat, Bakk.techn.
Matr.Nr. 0003520
Passettistrasse 91/8
1200 Wien

17.08.2016 _____

Kurzfassung

Diese Arbeit zeigt Möglichkeiten eines interdisziplinären Ansatzes auf, um die Spezifikation, die Evaluierung und den Einsatz von psychoanalytischen und neurowissenschaftlichen Annahmen zu unterstützen. Computer-Simulationen eines integrativen Modells der menschlichen Psyche helfen dabei die Trennung von individueller und sozialer, sowie emotionaler und kognitiver Entscheidungsfindung zu beseitigen, um ein ganzheitliches Erklärungsbild und detallierte Vorhersagen zur Verfügung stellen zu können. Die dabei erkannten Einflussfaktoren ermöglichen die Erstellung und Evaluierung von Handlungsempfehlungen, um kooperatives und umweltfreundliches Verhalten forcieren zu können. Die zentralen Fragen, die in dieser Arbeit für die Verfolgung dieses Ansatzes beantwortet werden, sind (1) methodischer, (2) theoretischer und (3) anwendungsorientierter Natur und können folgenderman zusammengefasst werden: (1) Wie soll ein Leitfaden zur interdisziplinären Zusammenarbeit gestaltet sein, damit Methoden aus der Computertechnik, Psychoanalyse, und Psychologie vereint werden, um (2) eine grundlegende mentale Architektur der menschlichen Informationsverarbeitung im sozialen Kontext zu entwickeln, die (3) in Simulationen getestet und zum Zweck der Erforschung von Einflussfaktoren kooperativen und umweltfreundlichen Verhaltens eingesetzt wird. Dieser Arbeit liegt die Hypothese zugrunde, dass diese Forschungsfragen durch eine Fall-geleitete inkrementelle Methodik mit iterativer Verfeinerung der Anforderungsanalyse und Spezifikation ihrer Annahmen auf unterschiedlichen Ebenen (Konzept, Modell, Implementierung) beantwortet werden können. Dabei wird angenommen, dass die funktionale Integration von psychoanalytischen Konzepten von Motivation (Triebe und internalisierte Normen) mit neurowissenschaftlichen Konzepten von Emotion und Gefühlen in ein grundlegendes Modell der menschlichen Informationsverarbeitung übersetzt werden kann. Dafür wird das Konzept eines schichten-basierten Zustandsindikator eingeführt, der unterschiedliche Prinzipien und Kriterien (unterschiedlichen aktivierte Bedürfnisse und Wahrnehmungen) berücksichtigt um mittels Bewertung von Zielen den Zustand des simulierten Menschens (Agenten) zu verbessern. Das resultierende SiMA-C Modell einer adaptiven Entscheidungsfindung stellt die Annahme dar, dass die Grundlage der menschlichen Informationsverarbeitung Datenaktivierungen, und -bewertungen sind, die nach Bedarf und Möglichkeit mittes Vermittlung zwischen Datenkonflikten und Datenauswertung erweitert werden. Insbesondere wird angenommen, dass diese Mechanismen mittels eines Zustandsindikators wirken und integriert werden können. Die Fall-geleitete Experimentierung und Erforschung des Modells in Simulationen zeigen, (1) dass eine konsistente interdisziplinäre Wissensübersetzung erfolgreich war, (2) dass die vorab spezifizierten detallierten Vorhersagen erfüllt wurden, (3) dass der Ansatz durch weitreichende Variablenmanipulationen zu ausgiebigen Verhaltenserkundungen geeignet ist und dabei die Forschungsmöglichkeiten der Psychologie und Neurowissenschaften erweitern kann. Insbesondere wird in Simulationen aufgezeigt, wie die Änderung von spezifizierten Determinatengruppen (das sind einzelne oder mehrere Parameter der Persönlichkeit, der Umwelt, und des internen Zustands) multi-kausale und facettenreiche Erklärungen aggressiven, kooperativen, ausweichenden und umweltfreundlichen Verhaltens bieten. Diese Ergebnisse eignen sich dafür Handlungsempfehlungen, die pro-soziales und umweltfreundliches Verhalten forcieren, zu formulieren.

Abstract

The thesis at hand demonstrates the power of an interdisciplinary methodology in specifying, evaluating, and applying psychological and neuroscientific assumptions. Computer simulations of an integrative model of human information processing support overcoming the separation between individual and social, and between emotional and cognitive decision making. The gained insights from such an explanation model are able to inform policies to enforce cooperative and environmentally friendly behavior. The key question to reach this objective is identified as (1 - methodical): What is a suitable guideline to combine methods from computer science, psychoanalysis, and psychology in interdisciplinary collaboration, for the sake of (2 - theoretical) developing a foundational mental architecture of human information processing in social context that (3 - applicative) is tested in simulations and applied to explore the impact factors of cooperative and environmentally friendly behavior? My working hypothesis aims at answering these questions using a case-driven methodology, which combines casuistics, software engineering, and simulation. These methods are used for an iterative refinement of requirements and specifications on the conceptual, model, and implementation level. The key assumption is that a functional integration of psychoanalytic concepts of motivations (drives and internalized norms) with neuroscientific models of emotion and feelings can be translated into a foundational model of human information processing. Therefore the concept of a layered state indicator is introduced, which the mental architecture uses to indicate (internally and externally) and enhance the state of unpleasure[1], pleasure, and internal conflicts. The underlying model of adaptive decision making conceptualizes memory-based data activation and valuations as the foundations of human information processing. These foundational functions are extended with functions of mediation in case of conflicts between contradicting valuations (i.e., unpleasure vs. pleasure) and with evaluation as a function of reasoning. Most importantly, these mechanisms are operationalized and integrated using the concept of a layered state indicator that uses different scopes and criteria in valuation, mediation, and evaluation. Case-driven model explorations and experimentations in simulations demonstrate the success of (1) a consistent interdisciplinary knowledge translation, (2) fulfilling the specified predictions, (3) the model's ability of extending the possibilities of the involved disciplines by wide variable manipulations. In particular, simulations show how changes in specified determinant groups (i.e., single or multiple parameters of personality, external environment, and internal state) provide multi-causal and multi-faceted explanations for cooperative, aggressive, evasive and environmentally friendly behavior. These explanations are able to inform policies for enforcing pro-social and environmentally friendly behavior.

[1] In the thesis at hand, the psychoanalytic term unpleasure instead of displeasure is used.

Acknowledgements

First of all I want to thank my first advisor Dietmar Dietrich for his support and continuous patience in emphasizing the principles of the SiMA approach, which he initiated and defended against all oppositions. He was always able to recognize distracting superficial questions, and led our research back to the chosen track by emphasizing the relevant questions. I also want to especially thank my second advisor, Cristiano Castelfranchi, whose work has accompanied and inspired me since my early master studies.

The inspiring atmosphere in the SiMA team was created by enthusiastic colleagues and intensive discussions with them. This includes fellow PhD students, (neuro)psychoanalysts, and psychologists. The opportunity to get a deep look into the research routine of other disciplines was very rich and inspiring. Stefan Kollmann with his openness for discussion and patience in fighting with (or rather against) the implementation for calibrating the model deserves to be specially mentioned.

After acknowledging the support of the academic atmosphere at the Institute of Computer Technology, TU Wien, I want to mention the support I received due to the atmosphere in my personal life. This starts with the direct and indirect support of my parents, who prepared the ground for choosing scientific interest as one *home* of my identity, although they did not get the chance for an academic education due to losing their *home*, which implicitly drove them to pass the importance of immaterial resources on their children. Thanks to my mother, who encouraged having an open mind and avoiding unnecessary categorization. Thanks to my father, who emphasized the role of knowledge and encouraged me to seek the answers to my questions on my own. Thanks to Tanja Fandler-Schieder, who accompanied an important part of this thesis and supported the work with her patience and tolerance.

I am also thankful for our democratic political and economic institutions, which provide indirect support by enabling the basic structures for society, and direct support by facilitating funding institutions such as the Austrian FFG (Forschungsförderungsgesellschaft), who also funded the CogMAS project, which enabled a part of the research behind the thesis at hand.

Deep thanks to Elisabeth Öfner for detailed proof-reading of my thesis and the emotional support in the final phase. Additional thanks to Swetlana Teutscher and Martin Fittner for additional proof-reading.

Table of Contents

Abbreviations

AI	Artificial Intelligence
ABM	Agent-Based Modeling
ABSS	Agent-Based Social Simulation
ACT-R	Adaptive Control of Thought-Rational
ANN	Artificial Neural Network
ARS	Artificial Recognition System
CE	Cognitive Emergence
CogMAS	Cognitive Multi-Agent System Supporting Marketing Strategies of Environmental Friendly Energy Products
DM	Drive Mesh
EC	Exemplary Case
EMA	Emotion and Adaptation
FATIMA	Fearnot AffecTIve Mind Architecture
fMRI	Functional magnetic resonance imaging
GPS	General Problem Solver
CLARION	Connectionist Learning with Adaptive Rule Induction Online
HCI	Human Computer Interaction
KORE	Kognitive Regelstrategieoptimierung zur Energieeffiziensteigerung in Gebäuden
ToM	Theory of Mind
ToMM	Theory of Mind Mechanism
NI	Neutralized Intensity
ORM	Object Relational Mapping
QoA	Quota of Affect
PP	Primary Process
RDF	Resource Description Framework
RDB	Relational Data Base
RREEMM	Resourceful-Restricted-Evaluating-Expecting-Maximizing-Man
SC	Simulation Case
SiMA	Simulation of the Mental Apparatus and Applications
SOAR	State Operator Apply Result
SoC	Separation of Concern
SC	Secondary Process
TP	Thing Presentation
TPM	Thing Presentation Mesh
UML	Unified Modeling Language

1 Introduction

Scientific questions and/or problems in the world are seldom separable by disciplines.

Unknown

Jedem Anfang wohnt ein Zauber inne ...

Hermann Hesse, 1877-1962

The development of an integrative functional model for the simulation of realistic human behavior in social context is still an open challenge. The thesis at hand sets out to provide a framework that enable gaining more understanding of how mechanisms of a simulated human's mental apparatus generate and determine decisions in a social context. This question will be tackled in an interdisciplinary fashion[1]. In particular the theories of other disciplines - mainly psychology (primarily psychoanalytic models) and neuroscience - will be used to tackle this question with a computational approach. Therefore open challenges in cognitive science and agent-based simulation will be addressed. Special focus is laid on a structured methodology in developing and evaluating an explanation model, in particular the special requirements of interdisciplinary collaboration (e.g., restricted accessibility of the human mind and complexity in its description) in bridging different disciplines and their methods.

1.1 Background Analysis

Various scientific disciplines aim at understanding how the mind works. Amongst others, psychology has brought important insights, but an integrated and holistic model of the human mind is still an open challenge (cf. (New73), (DSB+13a)). One hurdle is the insufficient collaboration between the involved disciplines in the humanities, social sciences, and natural sciences (Sun12, p. 6). A strict separation between the disciplines is recognized as an obstacle and is already decreasing, with the tendency to bridge different disciplines using (technical) natural scientific

[1]The thesis at hand is conducted in the context of the SiMA and CogMAS projects, where the author regularly worked with psychoanalysts, psychologists, and neuro-scientists.

methodologies. Such an approach requires intensive interdisciplinary collaboration and a dedicated interdisciplinary methodology bridging multiple required methods.

The thesis at hand aims to join computational cognitive science and agent-based simulations to bridge different disciplines (psychology, social sciences, neuroscience, computer science) in examining the individual mechanisms and impact factors of human behavior in a social context.

1.1.1 Computational Cognitive Science

The advent of information theory provided a new methodology for understanding how the mind works, which eventually led to the fields of Cognitive Science and Artificial Intelligence (AI). After fundamental research, the latter drifted off into solving narrow application-oriented problems and dismissed the original endeavor to take the human mind as an archetype. Furthermore, serious and regular interdisciplinary cooperation between the humanities, social sciences, and computer technology using a common framework has been neglected (Sun12, p. 6). The Computational approach to Cognitive Science has recently been used to simulate the mind, aiming to develop a computational framework model (Ber14, p. xxvii). Such synthetic approach can be regarded as a key method to understand the human mind (Bac11). It allows us to test our ideas of how the mind works by running them as computer simulations. Additionally, in an engineering sense, building a system requires understanding the system – and, of course building a system also deepens the understanding of it[2]. As common in engineering, this is done by testing if we develop the right model, and if we develop the model right. After understanding human information processing via simulation, we are also able to develop applications with human cognitive capabilities. This also emphasizes the need for fundamental research as a basis for applied research.

Another often neglected aspect is the development of a holistic and integrated functional model (instead of task or process models) (FMS07). Cognitive architectures are an important means for such an approach, representing a specification of a domain-independent structure that achieves the function of cognition (Sun06a, p. 33) (see Chapter 3.2.1). But over the last decades the focus in cognitive architectures lay on rational decision making and planning, with little consideration for affective processes. This follows the conventional separation of cognition and emotion in psychology, with focus on cognition, assumed to represent capabilities that distinguish human from other animals (as e.g. in (ABB+04)). However, the relevance of affective and unconscious processes is stressed, amongst others, through neuro-psychoanalysis, psychology and neuroscientific models (e.g., Bargh (BC99) and Damasio (Dam03)). Thus, to understand human decision making, we have to consider its basics, amongst others emotions. In this regard, we should rather aim at a *mental* architecture, representing all basic aspects of the human mind, instead of a *cognitive* architecture. This also would allow accepting different reasons for behavior in different personalities.

Additionally, conventional cognitive architectures and decision-making models are abstracted from the body. Following an embodied approach, human decision making is based on the interaction of the brain with the rest of the body (e.g. (NBW+05)). Another open challenge is to bridge social and psychological aspects of the human mind in cognitive architectures. Here the challenge is to find mental functions that generate, and hence explain, social phenomena and integrate them into a holistic cognitive (or rather mental, see above) architecture, which is the

[2] or even more extreme, following Feynman's popular quote "What I cannot create, I do not understand" (found written on his blackboard).

only tangible entity in a social system (Cas00). These mental functions are the mental mediators of social behavior.

These challenges can only be tackled using intensive interdisciplinary collaboration, which requires a shared methodology (see Chapter 2). In such an approach, psychology and neuroscience provide theories of the mind, and computer technology provides methods to develop and test a deterministic model of the human mind by using approaches from information theory, such as computer simulations, layered models, separation and modeling of data and functions, top-down design process, and software engineering. In summary, to express theories in a computerized form helps to specify, refine, and test them. Simulation can thus be used as a method of theory development, as a model building tool and as an evaluation tool. Such an interdisciplinary approach requires theories that are compatible with a computational methodology. In this regard causal functional models (that describe functions that generate behavior instead of behavior per se) are necessary.

Summarized, the consideration of the rules of the unconscious, affective processes, the subjectivity of human decision making, representation of personalities, embodiment, bridging social and psychological models, and interdisciplinary collaboration remain ongoing challenges in developing a holistic computational framework for Cognitive Science. When considering the mind as an information processing system, computer science provides expertise to examine and develop information processing systems, and other disciplines, such as psychoanalysis, neuroscience, and psychology, provide theories of how the mind works.

1.1.2 Agent-Based Social Simulations

The agent paradigm is used in different disciplines. According to the requirements of these disciplines, the definition of an agent is focused on different features of the agent paradigm. Most researchers use autonomy as one key feature to define the agent concept (Fra97). A widely used definition in AI describes an agent as a rational entity (which requires autonomy) that senses the environment and decides how to act upon it to reach its goals by maximizing its expected performance (RN03, p. 32ff.). The agent paradigm is also frequently used in computational social sciences, due to its suitability for a bottom-up approach to examining emergent phenomena in (artificial) societies (e.g. (Axe03)). In particular using agents enables the consideration of heterogeneity through individualization of societies, which supports bridging the gap between micro (agent) and macro (society) levels (Eps99). However, agents for social simulations are defined with different focus than AI. Although regarded as autonomous entities, an agent is defined through its interactions with other agents, and hence such systems are immanently multi-agent systems.

In general, simulations are useful for performing experiments that are not possible or too costly in the real world, building prediction models, discovering new relationships and principles, and for explaining phenomena. The modeling cycle consists of abstracting and formulating a model of the real system, simulating it, and interpreting the results. However, a simulation is not better than the assumptions built into it, and a computer can only do what it is programmed to do (Sim96). "Nevertheless, simulations can provide new knowledge: even when we have correct premises, it may be very difficult to discover what they imply. The more interesting question is whether simulations can help if we do not know very much about the natural laws that govern the behavior of the inner system" (Sim96). This is especially the case with the paradigm of agent-based social simulation (ABSS). Agent-based social simulations provide an approach for

interdisciplinary model building and evaluation (Eps99). In essence, ABSS explains a system's complexity at the macro level, e.g. social phenomena, with the frequent interaction of local components at the micro level, i.e. assumed and specified processes that generate the agent's behavior. These generative rules are usually formulated from other disciplines, which is termed as a novel kind of interdisciplinary collaboration (Eps99). Following a generative approach, ABSS is particularly suited to explain how a system's behavior is generated by underlying processes or mechanisms, with a special focus on the linkage between micro and macro level. In this regard, simulation is termed as a third way of performing science (Axe03), consisting of deductive and inductive means. On the one hand, instead of deriving causal relations from statistical analysis, assumptions are specified and tested based on empirical data. On the other hand ABSS also generates empirical data. Such generative explanations offered by ABSS are claimed to provide a deeper causal understanding of the phenomenon than do statistical explanations, which observe that in general, across a large number of empirical investigations, a particular regularity (e.g. correlation) is found (SC07). In addition to model development and evaluation, ABSS may guide empirical research by providing a theory-building tool (SC07). This is especially the case when relating micro and macro-level properties. Hence, ABSS also allows us to ask and test new questions and provides possible counter-intuitive insights (macro-surprises despite complete micro-level knowledge, e.g. the discovery of new relationships and principles (Eps08)).

When using ABSS, models may be tested at both the micro and macro levels (SC07), thereby implying two separate types of falsifiability: the first questions the validity of the underlying processes, while the second analyzes whether the generated behavior matches with the observed behavior, i.e. whether the model fits the empirical data. This happens precisely when the micro-specifications generate the macro behavior. However, empirical validation of ABSS is hard to achieve, as shown in an examination on a sample basis (DHK+07), where most ABSS that are not empirically validated are those that simulate the complex behavior of humans. This is mainly due to difficulties in sufficient acquisition of psychological and social data. But even in the case where empirical validation is hard to achieve, simulation of an agent-based model shows the plausibility of its assumptions, since it validates its consistency and to explain phenomena.

Besides sufficient empirical validation, another important open challenge in ABSS is the development of realistic detailed decision making models that go beyond the limitations of the classical rational choice approach, and hence continuing to bridge the micro-macro gap by bridging the socio-psychological gap.

1.1.3 Cognitive Computational Social Science

Despite some efforts in the field, social sciences have not yet met the challenge of incorporating a cognitive theory of social behavior as a research tool (Sun12). Computational social simulations represent a promising approach to tackle social research questions and bridge cognitive and social theories. However, the use of simplified decision models following a rational choice approach in simulations is often criticized due to their limitations in accounting for social phenomena (e.g. (JJ03), see Chapter 3.1.1). Merging psychology and social sciences using agent-based simulation (ABSS) is a promising approach to overcome these limitations, but early attempts often reduce decision making to single factors (see Chapter 3.1.2).

As common in modeling, ABSS (Agent-Based Social Simulations) strives to abstract and reduce information. The first step to tackle the micro-macro gap in ABSS is done abstractly (following a top-down approach in science) and has brought important methodological insights in showing how

the paradigm of agents - as heterogeneous individuals with frequent interactions - is able to tackle emergent social phenomena[3] (Eps99) (also see Chapter 3.1.2). Oversimplification in ABSS is increasingly criticized since their unrealistic assumptions impede their evaluation and application under real-world conditions (e.g., (Cas06), (Via12), (Sun06b)), (JJ03)). Additionally, in reducing social simulations on the interactional aspect, conventional approaches often neglect how social interactions are generated. These weak points are recently tackled by different approaches, which consider computational models of psychological processes that generate social interactions (e.g., (JJ03), (Sun06a)). Hence, in a top-down manner, the next iteration in computational social simulations has to be the further specification of simplified agents. Such psychologically plausible agents in ABSS can overcome the mentioned limitations and bridge the gap between psychological and sociological perspectives by providing a platform for their integration. It is particularly appropriate for (a) enabling model-development by integrating social and psychological variables into an agent, and (b) avoiding reduction to just one of these variable sets when explaining social phenomena. Of course, in the sense of Occam's-razor[4] simplification is favorable for validation, understanding and development - a case impressively emphasized with Braitenberg's vehicles (Bra84). However, it can be of advantage to follow a holistic approach first before we are able to know what to reduce and to parametrize such models realistically. In any case, the foundations of decision making have to be represented. With emotions being such a key mechanism to understand and predict social behavior, associated questions about empathy and mindreading are ongoing challenges in the field that explain social phenomena with mechanisms of the mind.

1.2 Motivation and Problem Statement

When aiming to simulate human decision making in social context for explanation and prediction, we have to understand how decisions are generated by demonstrating the impact and causal relationship of foundational variables in a holistic context. Conventional rule-based and rational choice approaches are limited in this regard. Even if necessary, it is not sufficient to use conventional cognitive architectures in social simulations to reach this objective, since they often lack the integration of emotion as the foundations of decision making and rarely consider social aspects (see Chapter 3). Nevertheless, the thesis at hand is not only motivated from the social perspective. Just as it is not sufficient to simulate social behavior without considering an individualistic aspect, the opposite is true as well. That is, when developing a computational model of human mental capabilities social aspects should be considered. When aiming to model the human mind, we cannot separate individual from social cognition: In general, the mental functions used for individual cognition are the ones used for social cognition. In this regard cognitive architectures and agent-based models often lack authenticity of functional equivalence (see Chapter 3).

Using such an approach is assumed to provide deeper explanations of (social) behavior and better predictions compared to conventional approaches. Accounting for the various parameters of human decision making realistically, enables a better connection to real world behavior and in turn a fine-grained model calibration with empirical data. Additionally, using behavioral realistic parameters enables the development of recommendations that fit and connect directly to the real world (without the need of interpretive mapping). Such an approach also allows for more powerful

[3] Much work in ABSS was concerned to show how the paradigm of agents is able to tackle emergent phenomena. Only quite recently the application of these methods with real-world data is focused.

[4] or the law of parsimony, "Among competing hypotheses, the one with the fewest assumptions should be selected." See https://en.wikipedia.org/wiki/Occams_razor

and flexible simulation experiments. And it enables experiments with variables that are difficult to manipulate in empirical studies, e.g. the external environment, or are not accessible at all, e.g. unconscious impact on decision making.

Deeper understanding of the human mind also enable technical systems that harness human capabilities. This is especially relevant for systems that require interaction capabilities or are embedded in social context.

1.2.1 Relevancy and Implications

Answering this thesis' question should address those open challenges and limitations in the field that are mentioned so far in the introduction. Summarized, these are the consideration of the rules of the unconscious, motivational and emotional processes, the subjectivity of human decision making, representation of personalities, bridging social and psychological models, and interdisciplinary collaboration. Using the presented methodology (see Chapter 2) enable developing and evaluating a computerized model that is able to generate and explain human behavior, which includes behavior often termed as 'irrational'[5]. Such simulations can be applied for the wide spectrum of human behavior that cannot be explained by classical rational choice models.

The aimed model should simulate human behavior realistically. The usage of a functional and holistic model of decision making enables experimenting how different factors influence an agent's behavior. We can observe how the change of factors such as personality, environment, memories, and the internal state of agents influence their social behavior, in particular how the dynamic interplay of these factors determines behavior. For the specific focus (of using a mental architecture to examine basic social interactions and environmentally-friendly behavior) this means to show for instance why 'Agent A' cooperates under specific conditions with 'Agent B' but not with 'Agent C', and when changing the conditions with none; or why 'Agent A' switches its energy provider when receiving 'message x' from 'Agent D', but not when being in another internal and/or external state or getting another message, possibly from another agent. Such simulations support exploring the conditions of aggressive, pro-social, and environmentally-friendly behavior by showing how a specific external environment determines behavior differently dependend on personality and the current internal state. Such fine-grained information allows for recommendations that ground policies to basic variables of human decision making.

Such simulation is not only a powerful tool to explore the range and combinations of determinants for specific behavior, but also a powerful prediction model, e.g., for decision support systems. This bridges basic research and application-oriented research. Hence, the presented model is not only a tool for different scientific disciplines, but a pre-requisite for various application tools, e.g. for developing and testing policies to distribute environmental-friendly products.

Various future application domains are enabled by the aimed model. Consider for instance crowd behavior, where such simulations enable the consideration of factors for preventing mass panic and herd behavior, by enabling prediction tools for supporting the decision of how to face mass panic as a policy maker. Such a tool also may be used to examine solidarity and how to enforce it. Of course, simulating realistic social behavior is also relevant for traffic simulations. Such models could also be used for simulation of cooperation in organizations (teamwork), such as group decision simulations for teamwork, through which it will be possible to create and test virtual

[5] From an external view behavior is sometimes regarded as irrational since it seems to harm the person. However, from a subjective point of view this does not necessarily have to be the case on the mental layer - in short-terms.

group decision scenarios. The role of emotions and empathy is also eminent in intelligent virtual agents, which are a promising approach to support professional training in various fields (e.g., physicians, policemen, and teachers) and where believable agents are required. Other concrete applications are found in consumer behavior (cf. the role of emotions and empathy in consumer decisions) and robotics. The former include simulations that enable decision support systems, such as supporting marketing decisions for disseminating ecological technologies. The latter supports modeling natural robot-human interaction with artificial companions, which requires the consideration of human-inspired mechanisms, such as empathy and emotions. Such systems are also able to consider the user's internal state and hence adapt better to the user and her goals. Thereby it provides lifelike interaction and enables adaptive, automatized and natural interaction.

In all these examples a holistic valuation model, in particular emotion and empathy, are key factors. Consider for instance, empathy and norms in teamwork, 'gut decisions' and social pressure in consumer decision, the significance of emotion contagion in crowd behavior, the influence of drivers' internal state in traffic, and the importance of a household robot to emphasize with humans for the sake of better understanding what they want. In the long term, the gained knowledge also can be used to enhance socio-technical systems, such as smart energy grids and machine interaction in automation, with human-like capabilities of social interactions and communication.

1.2.2 Excursus: Towards Non-Autistic Simulated Humans

Instead of specifying models of the human mind from other disciplines into a consistent and testable form, often special models that suffice computational criteria (such as efficiency) are developed. In this regard computer science stays behind its possibilities in contributing to understand the human mind.

When taking the human psyche as an archetype for a cognitive architecture, we have to find the psychological functions that generate human behavior. The significance of using plausible functions in developing a cognitive architecture can be shown by observing the behavior of humans that lack such functions. In the case of some social abilities this is the case with autistic people, who need a strict and inflexible scheme of rules to be able to interact with other people (BC12). Autism shows that a rigorous rule-based approach is misleading when simulating social interaction, and that our ability to interact with other people is based on emotion and empathy. That is, the ability to identify what someone else is thinking or feeling, and to respond to that with an appropriate emotion (BC12).

Hence, the difference in functions used by autistic people and non-autistic people exemplifies the difference between conventional approaches in agent-based social simulations or cognitive architectures and the approach of the thesis at hand. In particular, it shows that conventional approaches that do not consider the functionality of emotions are not able to sufficiently model human-like functionality. In this regard we can speak of autistic agents.

1.2.3 Scientific Questions and Problem Analysis

Following a generative-functional and holistic approach (see Chapter 2), the general objective that guides the thesis at hand is to find computational representations of mental functions that

generate behavior in social context and data that determines it. The general methodical question of the thesis at hand, then, is ...

... how should we develop and test an integrative model of theories from disciplines involved with the human mind, and apply it as an explanation and prediction model in simulations?

Tackling this foundational question includes an analysis of how human psychic functions generate (social) behavior and how these functions may be represented in a computational model. Hence, with the SiMA research program (see Chapter 3.2.6) as a frame, the general theoretical question is ...

... how does a foundational mental architecture generate human behavior under consideration of the social context?

Using a top-down problem analysis this general question (the 'Top') must be first addressed abstractly, before reaching specific sub-questions. A sketch of this analysis is shown in Figure 3.14.

With the premise of an interdisciplinary approach, where the task of computer science is to provide a shared methodology and appropriate tools in tackling the research question, and other disciplines provide the theories, one key task in answering this thesis' question is to develop a methodological framework that enables answering the research question by following a structured procedure and criteria (see Chapter 2). This includes answering such questions as: How to translate knowledge from psychology and social sciences into deterministic models for a human-like agent? How to specify abstract models from other disciplines for a computational model? In doing so, how to merge methodologies from different disciplines? How to evaluate the resulting model and the underlying assumptions? In fact, these pose the biggest challenges in the thesis at hand. Overall, the triggering questions come from psychology and social science, e.g. which factors determine cooperative or environmentally friendly behavior, but the main question addressed by the thesis at hand is how to use computational methods to support answering these questions.

The problem analysis sketched in Fig. 3.14 can be extended by further questions, such as: How to generate possible actions and determine their relevance, given an agent's (possibly conflicting) different demands? How to adapt this decision on its internal state and external environment? After developing first theories about the general research questions, further sub-questions include first assumptions. For instance, with the assumption that a generic model of data activation and valuation is able to account for goal generation and selection, the question arise: How do agents use activation and valuation mechanisms, inspired by motivation and emotion, to decide if and how to interact with another agent, or to act in an environmentally friendly way, respectively? How does the perception of other agents influence the subject's motivation and valuation? In the remainder of this thesis such sub-questions are implicitly stated and explicitly tackled in the according chapter.

The thesis at hand aims to set the ground for elaborated simulations of social behavior. Thus, it only considers foundation of social behavior, and neglects elaborated concepts for social interaction, such as explicit communication between agents (e.g. verbal communication). Before such behavior can be simulated by a generative-functional approach, its pre-requisites must be set by developing the foundations of decision making.

The aimed mental architecture enables flexible simulation experiments under different conditions. On the one hand, it enables the analysis of (the interplay of) internal and external conditions that

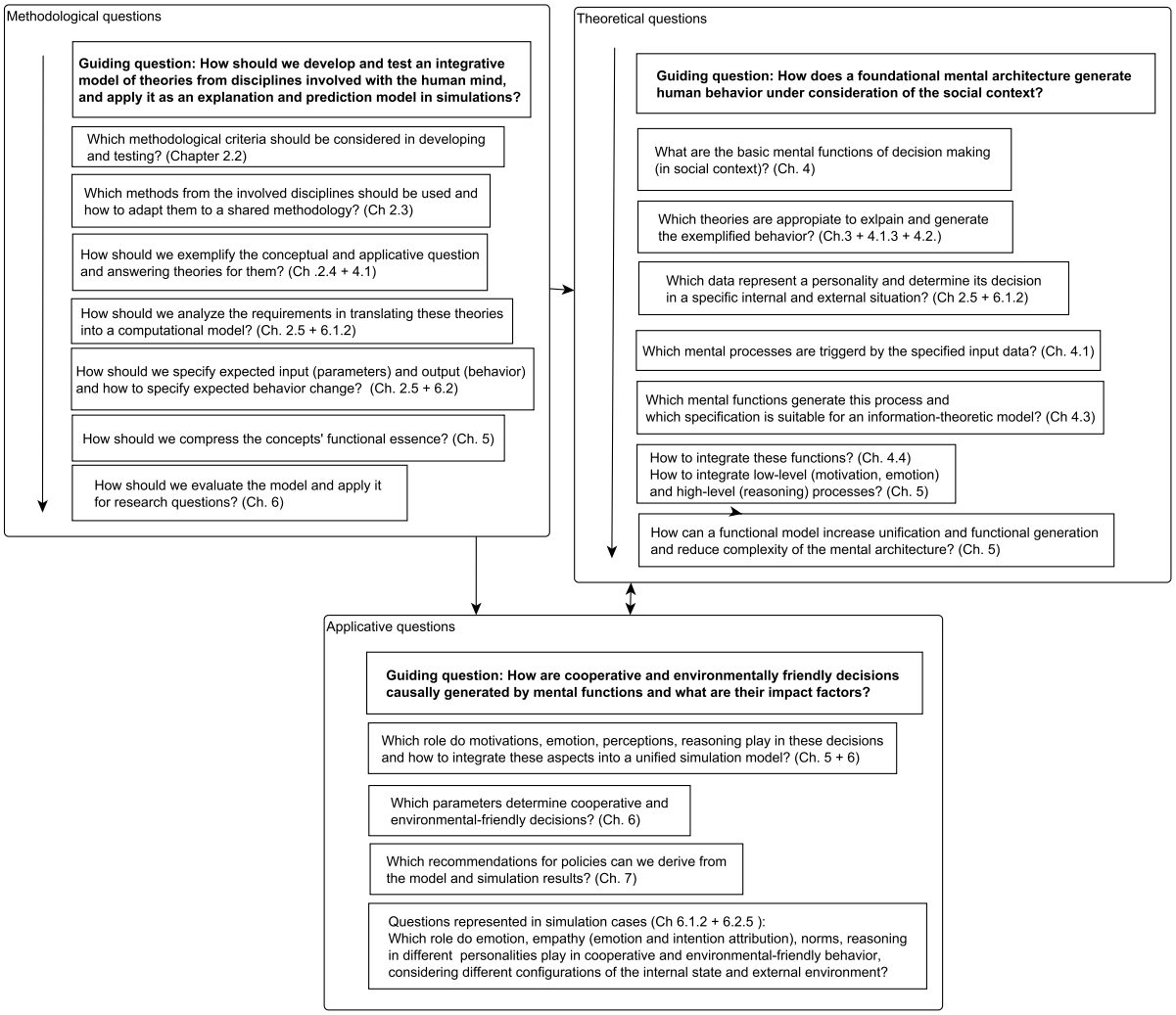

Figure 1.1: Top-down problem analysis on a methodological, theoretical, and applicative level.

lead humans to act in a specific way. This also supports the demonstration of the dependence between decision variables. On the other hand, it enables the recognition of universal principles, which are personality-independent, in a specific domain, e.g. cooperative or environmentally friendly behavior.

The described general questions are expressed in two domains: social interactions in artificial life simulations, especially cooperative simulations[6] and environmentally friendly behavior[7]. In particular, the aimed mental architecture should support answering the question of why humans cooperate and why they act environmentally friendly, i.e., why agents cooperate with other agents and with nature.

To examine cooperative behavior a paradigmatic scenario of two agents and a food source is used. This enables determining the conditions in which different agent-types share the food source, offer it to the other agent, eat it alone, avoid the situation and go away, or even fight for the food

[6]This part of the work is done in the SiMA-Basic project of the Institute of Computer Technology.

[7]This part of the work is conducted in the FFG-funded CogMAS-project, a cooperation between the Institute of Computer Technology (TU Wien) and the Institute for Marketing and Consumer Research (WU Wien).

source. From this exemplary simulation we are able to derive general insights about the conditions that support aggressive or cooperative behavior and possibly could derive recommendations.

Environmentally friendly behavior is examined with the exemplary case of switching to a green electricity provider after an agent gets prompted in social media. Besides the scientific insights about psychological mechanisms and impact factors of environmentally friendly behavior, the results serve the development of a decision support tool for testing distribution-policies of environmentally friendly electricity products. The social media scenario supports exploring the impact of social pressure and the limits of using messages from peers to motivate behavior. As generally in the followed approach, the impact of the external environment is observed in different internal states and personalities to get fine-grained insights about the impact of variables in context.

An exploration and analysis of these two paradigmatic cases will reveal further concrete requirements and assumptions. Given this thesis' approach the functions that generate behavior for both cases are assumed to be the same, with the only requirement of different parametrization.

1.3 Working Hypothesis

The top-most assumptions in the thesis at hand can be described as follows: The methodology (described in the next chapter) is assumed to support and guide knowledge transition into a generic computational model, which represents the top-most hypothesis of the thesis at hand. We assume that - using the presented methodology - the selected theories are transformable into an explanation and prediction model that is testable in simulations. The developed methodology is assumed to enable applying a generic model for a wide range of conditions in one application domain. An additional assumption about the advances of applying the methodological criteria at hand, is about the applicability of the specified concepts in different domains. In particular, I assume that the resulting model is capable to generate and examine the target behavior - social interactions and environmentally friendly decisions. The key hypothesis in the resulting model is that valuation mechanisms and their pre (activation) and post-processing (mediation and evaluation) are the foundations of decision making with consideration of social context and able to account for the target behavior. Furthermore, they are assumed to be the basis for a generic simulation model that is able to examine the reasons for behavior in different domains, be it cooperative behavior or environmentally friendly behavior.

1.4 Relation to Other Publications

Parts of the thesis are based on the author's publications (see literature list at the end of this thesis). This concerns only parts of these publications that are written by the author of the thesis at hand. Chapter 2.3 is partly published in (Sch16a). Chapter 5 is partly published in (Sch16b).

2 Approach and Methodology

Ideas such as absolute correctness, absolute precision, final truth etc. are fantasies that should not be allowed in any science ... since the faith in one single truth and being their owner is the deepest root of all viciousness in the world.

[Original in german: "Ich glaube, daß Ideen wie absolute Richtigkeit, absolute Genauigkeit, endgültige Wahrheit etc. Hirngespinste sind, die in keiner Wissenschaft zugelassen werden sollten. ... Ist doch der Glaube an eine einzige Wahrheit und deren Besitzer zu sein, die tiefste Wurzel allen Übels der Welt"]

Max Born, 1882-1970, "Von der Verantwortung des Naturwissenschaftlers".

Science is in the technique. All else is commentary.

Allen Newell, 1927-1992, Newell's last lecture (http://act-r.psy.cmu.edu/misc/newellclip.mpg)

To reach the objective of the thesis at hand, an integrated interdisciplinary methodology for developing and evaluating the model must be specified. This comprises methods for theory, model, and software development and evaluation, with the consideration of their relation. Different than in other areas, in the context of cognitive architectures currently no established scientific methodology of development and evaluation exists. Especially the methodology of interdisciplinary knowledge translation and model evaluation is debated on and their scientific nature is questioned (Coo07). With the claim to develop virtual humans, how can we validate that our model represents human functioning? More precisely: How can we translate knowledge from psychology and social sciences into deterministic models for a human-like agent in an interdisciplinary fashion? How can we specify abstract models from other disciplines appropriately? How can we merge methodologies from different disciplines? How can we evaluate the resulting model, and what does that mean for the validation of underlying theories?

The presented methodology, termed case-driven agent-based simulation (SD14), merges methods from computer science and psychoanalysis, in particular casuistry, empirical studies, software engineering, requirement analysis, agent-based modeling, and computer simulation. It bridges disciplines with different methods and provides a shared methodology. Furthermore, it considers key challenges of the research domain. These are in particular the restricted accessibility of the

human mind, interdisciplinary understanding, the complexity in generating human behavior and its relation to social phenomena.

There are various ways to get information about the mind's functioning, but we have to consider if theories are relevant for the objective at hand and if they fit the criteria of the SiMA research program. However, usually theories from other disciplines are not directly useable for developing a computational model and need interpretation and specification. Additionally, different assumptions derived from these theories may be inconsistent. In any case cooperation with experts from the respective field to interpret and translate these knowledge for our purpose is required. Therefore a shared tool is necessary. However, in such collaboration between researcher from different disciplines, often using different vocabulary, common understanding can be difficult. This is especially the case in regular intensive collaboration, with a complex system as the mind as the research topic (see Chapter 3.1.2 for the issue of complexity)[1].

2.1 Appropriate Methods from Science and Engineering

In a nutshell science can be regarded as a method aiming to explain[2] phenomena for repeatable predictions (Det07). A key driver of science is asking good [3] research questions and to doubt given answers. Science, then, is concerned with the search after insights and knowledge in answering these questions. Opposed to other approaches, such as religion, this knowledge comprises of opinions (or statements) that are established and justified using rigorous, critical, and consistent methods (Det07). These methods traditionally are distinguished in deductive and inductive. The former tries to deduce facts from other facts and axioms (assumed facts), the latter aims to find facts from a set of empirical instances. Even if applying these methods is a necessary condition for science, it is not sufficient to demarcate it from non-science, since no *universal* statement can be proven with a *finite* number of (empirical) instances[4]. Therefore Popper proposes the falsifiability of assumptions (hypotheses) as the demarcation-criterion to non-science (Pop59). The scientific method, then, is concerned with testing a hypothesis' prediction in experiments. If empirical observations falsified the hypothesis, it should be dismissed. Otherwise, a set of consistent and coherent hypotheses that had stand the test of falsification can be eventually compiled to a theory, which in turn can be applied (e.g., instantiated in a model) or tested further (McC02). Thus, a theory is regarded as scientific, if it is principally falsifi*able*. Statements that claim to have universal exploratory power, and hence are not falsifiable, are then regarded as myths (Pop59). However, falsification as a method is rarely applied rigorously. That is, rarely are experiments conducted with the aim to falsify assumptions. Hence, in such approaches 'scientific' is attributed to a theory, not to a method.

Such extreme approach to science (to dismiss a hypothesis in the case of false predictions) is well applicable in many fields, in particular in those where 'atomic' hypotheses or small theories are

[1] The complexity issue is not only valid on the neural level, but also for the abstract mental level, since only the interplay of multiple mental factors determine behavior.

[2] Opposed to that, a scientific law is just a generalized description of observations (Hon95).

[3] To define what 'good' in this context means is subjective. Typical candidates for basic research are those questions that promise to have an impact on the field by contributing new insights, without considering their relevance for solving concrete problems in the world - which is rather a metric for research questions in application-oriented science. In any case questions may imply established theories or even hypotheses.

[4] This logical inconsistency is also known as the Hume-Problem (http://plato.stanford.edu/entries/pseudo-science/).

tested [5]. But for other scientific approaches Popper's methodology is not appropriate, particularly for holistic theories, such as cognitive architectures, which hence are sometimes regarded as unscientific (Coo07). But as mentioned, the extreme prescription of such approach is rarely applied. First, experiments cannot be conducted without commitment to a theory. Second, developing experiments with the goal of *falsifying* a theory is the exception in science. And even theories that fail to predict every observation are rarely dismissed (e.g., because the reasons for falsification are not clearly attributable to the theory, given necessary auxiliary hypotheses), but developed further to account for the negative test result (Lak70). In this light Popper's methodology is called naive falsifictionism by Kuhn and Lakatos (Lak70). Instead of classifying theories as scientific, methods should be classified as scientific (or not). Instead of theories, research programs should be the topic of science, with the demarcation-criterion of progressive theory development. In particular, a research program is scientific, if it is cumulative, i.e., if it progresses in testing its predictions and adapts its assumptions based on the test results. To be able to do that, theories should distinguish non-falsifiable core assumptions from testable peripheral assumptions, which should be incorporated into the core assumptions after they are corroborated in experiments (Lak70). For theory development to be progressive, a plan (as 'positive heuristics' [6]) that serves as a guideline of how to develop and test peripheral assumptions has to be specified in advance.

In this regard, cognitive architectures injunctively include universal statements and are not testable with Popper's criterion: They are only able to make predictions as models in an applied context, which requires task knowledge, which in turn represents a separate theory (Coo07). In testing such parametrized architectures, prediction failures are difficult to be attributed to the cognitive architecture or the task model. Additionally, the existence of theory-independent parameters (with the opportunity of parameter fitting) and implementation assumptions further impede falsification. Hence, a cognitive architecture can only be tested in the context of its application. To falsify the cognitive architecture per se would require tests with all possible task models, implementations, and parameter-settings. Thus, cognitive architectures should be developed as Lakatosian research programs (Coo07). Tests of cognitive architectures have to be specified to corroborate or falsify single peripheral assumptions. Additionally, cognitive architecture and task models should be considered as separated research programs to avoid attributing errors preferably to task models (parameter fitting) and hence impede the development of the cognitive architecture (Coo07). In general experiments should be conducted with a limited possibility of varying parameter and implementation details, otherwise cross-validation, where a model is tested against previously unknown data, should be conducted.

A central question, then, is how tests should be conducted and integrated in developing a theory. The overall objective is to assess how good a theory represents a targeted real process (which is particularly relevant for models, guided by theories in abstracting the target process). Such objective is usually regarded as validation (OSFB94). However, such usage of the concept of validation is misleading, since it is used as a synonym for verification, with its impossibility to determine the truthfulness of a model (since natural systems are not closed and an complete description not possible) (OSFB94). Not truth but valid models establish *legitimacy* in the sense that "A model that does not contain known or detectable flaws and is internally consistent

[5] The dominance of Popper's approach led to the decline of holistic theories that try to capture assumed relevant factors of a research topic. This led to fragmentation of knowledge in many scientific disciplines. For instance, contemporary psychology is mainly involved with testing assumed causal relations between a small set of variables. Criticism of this approach led to the research program of cognitive architectures (New73).

[6] Lakatos mentions Newton's iterative procedure as an example (Coo07)(Lak70)

can be said to be valid" (OSFB94, p. 642). Instead, to test the model, the notion of *partial* confirmation, which support the probability of a model - and hence corroborates the model - is proposed (OSFB94).

Even given Oreskes' interpretation of what validation means, a model's legitimacy has to be evaluated regarding its purpose. For instance, prediction models must be assessed under other criteria than explanation models. Of course, explanation generally entails prediction (and vice versa), but for instance, for the purpose of explanation parsimonious models are preferable (GT05, p. 20). In this regard, based on their purpose, (GT05, pp. 40ff.) distinguishes (1) abstract, (2) middle-range model, and (3) facsimile models, which also require different validation methods. Only in facsimile models empirical validating should be aimed, under consideration of the limits (e.g. biases by incomplete data). However, a validated facsimile model can only represent a valid candidate model of the targeted real process, which emphasizes again the impossibility of scientific verification.

Validation against a model's purpose is especially important for applied models, which aim for functional representation. That is, they do not aim to represent a target per se, but its functionality. Such approach is used in engineering, where the main objective is to create artifacts (given constrained resources) and test how good they work in solving concrete problems in the world. This entails verification against an artifact' specifications, which is possible since such artifacts can be described as closed systems (OSFB94)[7]. A typical engineering procedure is to identify and define a problem, analyze the problem's requirements, define the solution's criteria[8], formulate specifications based on requirements and criteria, search for possible and innovative solutions (i.e., inventions), evaluate them against the criteria, design and implement the assumed best solution in an artifact, and test it. Verification, then, is concerned with testing if the artifact fulfills the specifications, and validation is concerned if the artifact behaves as expected in the application domain and solves the problem as expected [9]). This typical engineering procedure follows the criteria of the scientific method and uses knowledge gained by science in building artifacts.

In general science not only benefits in using the artifacts of engineering, but can also benefit from its methods. How to build a computational model based on hypotheses is an interdisciplinary engineering task, using requirements engineering and technical methods, such as the methods of computational model development (e.g., software engineering). Overall, the scientific method provides a framework and the engineering method concretizes it with suitable methods for model development and testing.

Using models as a support or means to solve problems, computer simulations follow a functional approach in testing them. That is, "a model should be developed for a specific purpose (or application) and its validity determined with respect of that purpose" (Sar13, p. 12). In this context model verification ensures "... that the computer program of the computerized model and its implementation are correct" (Sar13, p. 12) and model validation is concerned with the

[7] "The assumption of verification is that everything can be completely dened in a formal manner whereas in validation this closed world assumption does not hold" (LHU09, p.161).

[8] Criteria concern the solution and establish a framework to achieve it. Opposed to requirements, criteria do not have to be derived from the problem. However, best-practices and experience may show that the consideration of specific criteria for a problem domain is necessary (in the long-term).

[9] The IEEE 1012 standard states (IEE12)(LHU09): "The validation process provides evidence whether the software and its associated products and processes 1) Satisfy system requirements allocated to software at the end of each life cycle activity 2) Solve the right problem (e. g., correctly model physical laws, implement business rules, use the proper system assumptions) 3) Satisfy intended use and user needs".

"substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" (Sar13, p. 12).

Given a simplified overview of the development process in a computational approach (see Figure 2.1) validation and verification can be related to the development process as follows (Sar13): Conceptual model validation is concerned with guaranteeing the correctness of the assumptions behind the conceptual model, given the purpose of the model. Computerized model verification assures that the implementation is correct, i.e., follows the specifications of the conceptual model. Operational validation determines that the model's behavior has a satisfactory range of accuracy for the intended purpose in an application domain. Data validity aims to guarantee the correctness of data necessary for model building, evaluation, and experimentation (Sar13). As these processes are partly dependent on each other, they should be conducted sequentially (first conceptual validation, then implementation verification, and last operational validation, with data validation done in every step) (Sar13).

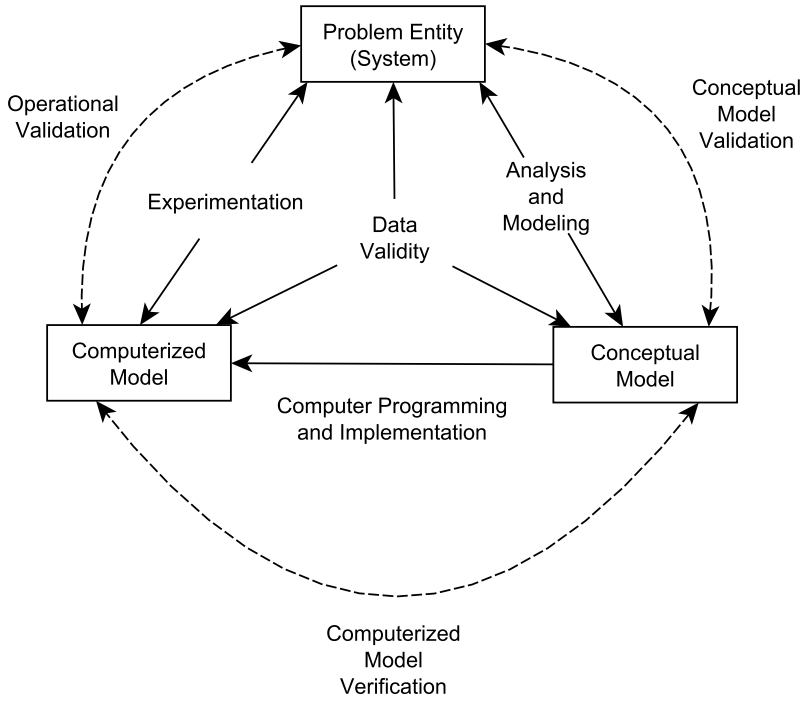

Figure 2.1: Computational Modeling Process (Sar13). A conceptual model is supposed to represent the target system. An computer implementation of this model can then be used for (virtual) experimentation. Each of theses steps require a (different) validation mechanism.

Thus the objective of conceptual validation is to test the assumption's logical structure and consistency and to test if the degree of detail is sufficient for the model's intended purpose[10]. From the various validation techniques that are used in (Sar13), for gaining conceptual validation, primarily face validation, structured walkthroughs, and traces are appropriate to use (Sar13), which all include a review by peers and experts of the respective field.

The suitability of validation techniques for operational validation is dependent on the system's (or problem entity's) observability, i.e., the possibility to collect data about its behavior. If

[10]Conceptual validation follows Oreskes' definition of valdiation (OSFB94) (see above).

this is the case, typically used methods are comparison of the model's behavior to the system's behavior and parameter variability–sensitivity. Concrete techniques for the former are (in their prioritized order) comparison of following entities between the real system and its model: (1) means, variance, distribution, or time-series of the output variable for specified experimental conditions; (2) confidence intervals of (1); (3) graphs" (Sar13, p. 19). If the system is not observable, domain experts should asses the reasonability of the model's behavior and parameter variability–sensitivity (Sar13).

2.2 Methodological Criteria

Criteria pose requirements on the methods for achieving the solution. Additionally they pose requirements on the solution that are not directly derived from the research question, but from the research program's approach and best-practices in following it.

Posing these criteria is also an implicit working hypothesis, i.e., assuming that meeting these criteria and following its principles support the objective of this thesis in developing foundational functions that generate and determine social behavior. Additionally, these criteria support consistency and compliance in knowledge translation and thereby eases interdisciplinary collaboration.

2.2.1 Conceptual Criteria

Conceptual criteria are central, since they implicitly pose theoretical assumptions. In this regard it is particularly important to justify and demonstrate why they are chosen to tackle the challenges of the research topic (e.g., restricted accessibility, complexity, interdisciplinary collaboration, knowledge translation). In a top-down approach these criteria are especially relevant for theory-formulation, specification, and modeling.

These criteria can be regarded as different aspects in approaching the research topic.

Interdisciplinary Aspect
Computer scientists build and use information processing systems. Considering the mind as such, computer science can be regarded as an appropriate discipline to explore the mind's functionality. This consists of specifying psychological theories into computational models and applying computational methods to evaluate these models. Although computational methods support the corroboration of assumptions, the used theories nevertheless should provide evidences of their plausibility.

Such approach requires intensive and regular interdisciplinary collaboration using a common platform. Psychological theories must be appropriate for the interdisciplinary collaboration with computer science and have to meet the conceptual criteria of this thesis. But the theory must be specific enough - without the need for too many auxiliary assumptions - to provide starting points, and also consistent enough to develop a non-contradictory model. This is particularly the case when integrating different theories, which is necessary when aiming to develop a holistic model of the human mind. Such consistent integration is supported by a computational model. However, these factors are only assessable in the process of developing the model.

In computer science typically task models are developed. That is, given a specific task, an algorithm is developed that fulfills this task. The only criteria for such algorithms are computational,

e.g., efficiency, complete rationality. But when using computational means to model the mind, other criteria must be met. Most importantly, the developed algorithms should be valid functional representations of the mind. For instance, formal theories of logics should not be implied generally for the functionality of the mind. Psychological theories (psychoanalysis included) provide concepts of how the mind works. Computer scientists should only specify and transform them into a computational model, using established principles, such as layered models, and the separation of function and data. In any case, the point of departure must be psychological theories of the mind, not computational algorithms. Only the second step is to find an appropriate algorithms that is a valid functional representation of a previously specified psychological function.

Even if the involved computer scientists have to understand the psychological concepts to translate them into a computational model, especially in the beginning respective researcher from other disciplines have to be integrated in the process in a regular manner. When aiming for a holistic approach (see below) the required broad knowledge of a discipline cannot be replaces by an intensive literature research by a computer scientist. Integration of researchers from the respective fields should also be considered in the ongoing review process - from the selection of theories to their translation into a computational model.

Following such an approach, aspects of the human mind that are often neglected in computational approaches are considered. This includes the relevancy of unconscious processes, embodiment, and emotion.

Holistic-Unified Aspect
To be able to analyze the unknown importance of various factors of a phenomenon, all assumed basic factors must be considered at first. The more uncertainty about the relevancy of factors, the more important is a holistic model. This is especially the case in basic research, where essential questions are still open. Such approach is better able to demonstrate the relevancy of factors and how they interact in generating behavior. Computational means enable following such an complex and resourceful approach. After an analysis, a holistic model, then, can be reduced for a more adaptive application.

A paradigmatic example is the omission of emotion in exploring decision making. The historic formation of two (contradicting) concepts and principles to explain why people behave as they do, namely thinking and feeling, with a more or less sharp separation between them was strengthened in Enlightenment, aiming to explain secular affairs by rational means. Until recently emotion was excluded, since it was not considered to be comprehensible by rational scientific means. The focus on logics in modeling human decision making, also in classic AI, is one expression of this development. However, following an evolutionary approach to the mechanisms that can be described as emotional or affective, they have to be regarded as the foundations of the human mind, since they are phylogenetically older. In this regard, an evolutionary development entails a hierarchical relationship between emotion and reason. This is also a basic principle in neuroscience: new brain functions must be integrated with old functions - on which they are based - and cannot replace them.

Using a framework theory, which enables the consistent integration of different factors and aspects, supports building a holistic model. One of the few psychological theories with such features is Freudian metapsychology (see Chapter 4.2.1). Such abstract framework also facilitates interoperability between different theories. A conceptual framework also enables to consider the holistic functionality of the mind. On the one hand, this allows an abstract perspective on holistic aspects of the function model, which can later be mapped to concrete aspects. On the other hand, it also considers different aspects of specific functions and the interplay with other functions, which

enforces interoperability between different (psychological) concepts. Hence, a holistic approach entails an integrated approach. In this regard a mental function is not approached isolated, but integrated into the holistic framework.

Aiming functional equivalence to the human mind, the model in turn should be simplified to a functionally equivalent model, which lead to a more unified model. This iterative process is again supported by using a holistic framework. It can be even argued that in a domain where the building blocks are unknown and aimed for, we need to first consider a holistic approach to be able to analyze and recognize the relevance of single factors (in simulation). The neglect of a holistic approach - when aiming for an unified model - has the tendency to lead to a (possibly inconsistent) patchwork model.

Generative-Functional Aspect
The premise of developing a functional equivalent model of the mind can be regarded a functionalist view on the mind (e.g. (BF72)), instead of a representational view. That is, the mind can only be described by its functional features. A similar concept is money, which only can be described by its functional features, but not by its physical features. However, opposed to functionalism the model should aim for a valid representation of the mind's information-processing functions, which is used in a computational sense (derived from the mathematical notion of function) and should not be confused with purpose. In this regard one can summarize that data serves a function, which serves a purpose.

In a conceptual sense such functional approach aims to develop a causal model. In the context of a holistic approach this requires the demonstration of a causal chain[11], which - based on axioms - comprehensibly explains how a mental function's behavior can be *generated* from a specified input data *spectrum*. Only basic axioms or statements upon them are allowed to generate the output. The behavior, then, is determined by (the change of) the function's parameters, not by (a change) of the function description. Such approach aims to find a candidate that explains a functionality, instead of describing it. An example is the functionality of motivation, which can be explained by drive functions. However, this example also shows the need to integrate and unify functions, since more concepts are required to explain the functionality of motivation holistically, e.g., norms. The next step would be to analyze the common functionality (e.g., of drives and norms) and integrate them into a unified function.

The usage of a descriptive model to gain such explanatory functional specification can be necessary initially. However, grasping a subject matter in a descriptive way (e.g., by categories and taxonomies), should only be a first step, since it is not able to provide the insights and applicability of an explanatory functional model. Overall, the aim should be to find a function model, with a behavioral model being only a mock-up possibility as a first approach.

The recognition of function demonstrates utility and meaning, i.e., its semantics. This also complies with a subjective approach and the principle of evolution: From an evolutionary view something only 'makes sense' if its function brings utility to the subject, i.e. it has a purpose for the subject. That is, in an evolutionary sense, the concept of semantics refers to functional utility for an subject.

Such a unified generative-functional approach not only provides a candidate explanation and hence drives theory development, but also enables finding simple patterns that explain complex systems by demonstrating how simple functions are able to generate complex behavior. Also, in a technical sense, such models are generic and flexibly applicable in different domains.

[11] A holistic approach demonstrates not only the causal relation between single variables, but also the interplay of all involved variables in a causal chain.

2.2.2 Deduced Criteria

The described conceptual criteria imply some other criteria that will be considered explicitly.

Following a generative-functional approach to explain social behavior implies following an approach of methodical individualism, with the claim that all social phenomena are the product of mental phenomena (Via00). This implies that social phenomena can only be explained with a bottom-up approach. Of course, a top-down impact of institutions and policies have to be considered. However, to consider social aspects, such as institutions, their impact has to be explained through the agent's mind (Cas00). This does not necessarily mean that social factors have to be related to self-centered factors, since the association of self to others may be strong or even (subjectively) identical[12].

Thus on the social level a bottom-up approach should be followed, but on an individual level a top-down approach. However, the (explanatory) gap between the generative assumptions (i.e., assumptions about the mind) and the emergent phenomenon (i.e., social behavior) on the social level is much smaller than the gap between the generative assumptions (i.e., assumptions about the neurons) and the emergent phenomenon (i.e., decision) on the individual level. Hence, to analyze the impact factors and identify the causal chain, on the individual level a top-down approach is more feasible and appropriate.

To build a holistic functional model a top-down process have to be followed. Starting from the most abstract theories and concept, they are specified into an implementation model. Only in this way a consistent, unified, holistic model can be guaranteed.

In such approach the relation between the used theories and their translation into computation models and implementation have to be coherent. The equivalence of theories, model, and implementation have to be tested in interdisciplinary reviews and specifications of test-cases.

Developing a unified, integrated model in a functionalist approach implies the requirement for a parsimonious model. With the objective of functional equivalency, the simplest model should be developed in an iterative manner. This not only serves understanding, but also implementation and application.

The interdisciplinary aspect require the consideration of the crucial role of the unconscious. In humans two qualities of information processes can be distinguished : unconscious and conscious (e.g. (BC99)). These distinction is often neglected in human-like agents although it is claimed (e.g., (BC99), (Kah13, p. 4)) that the majority of mental processes occur unconsciously.

In general an interdisciplinary approach entails developing an autonomous and adaptive agent. This is also entailed by a functional-generative approach, since an agent should have the functionality to generate its own agenda and adapt on the inner and outer environment its decision making.

2.2.3 Implementation Criteria

The described criteria directly or indirectly affect how the resulting model should be implemented. Additionally, best-practices of software engineering should be considered. This includes the separation of the model's description and its implementation, since they have to meet different criteria.

[12]Hence, this approach should not be mistaken for a objectively self-centered approach. In this regard, also, it has to be considered that the idea of independent individuals is quite new, and ideologically loaded.

The model have to be an parsimonious functional integration of different psychological theories, with the possible intermediate step of a process-flow description (see below). For instance, the model should be unified, but its implementation should be modular (following an established software pattern). Also the implementation has to consider the extensibility and maintainability of the software design. Additional implementation criteria include the robustness and balance between a generic and concrete software design. This also supports reducing the implementation complexity and adapting the implementation complexity on the model.

Where possible, established frameworks and software should be used, to support focusing on finding an appropriate implementation for the model, instead of looking for technical solutions for already solved problems.

In implementing the model, an iterative top-down design process as sketched in Figure 2.4 should be followed. This complies with the holistic approach and additionally serves incremental and agile development that adapts on ongoing requirements analysis and is able to consider early feedback from simulations and tests; hence, it recognizes pitfalls and bottlenecks early and accounts for them timely.

2.3 Case-Driven Simulation

A research project is typically embedded in a research program, similarly a hypothesis usually has (implicit) premises. In our case this is the development of a mental architecture for application in Cognitive Science and Artificial Intelligence in the context of the SiMA project (Simulation of Mental Apparatus and Applications, see Chapter 3.2.6). The thesis at hand is mainly concerned with the first kind of application, i.e. to provide tools for psychology and social sciences to tackle questions with an social impact. Examples that will be shown are 'What are the mechanisms and impact factors for cooperation and environmentally friendly behavior'. Support in answering this question enables the specification of policies to increase cooperative and environmentally friendly behavior. How to adapt a existing mental architecture and develop a model to tackle this questions is the main research question of this thesis. Hence, a research program together with a problem domain form the context for a concrete research questions. But they also pose first constraints and assumptions (e.g., derived from the SiMA principles), which can be considered as frameworks assumptions, since all other questions and answers need to be embedded in them.

To tackle the research question a methodology is developed that considers the introductory mentioned challenges of our interdisciplinary research topic. Five iterative and incremental steps can be distinguished, each guided by questions that have to be answered for the sake of tackling the research question. An overview is given in Figure 2.2, the details are described in the following sections and sketched in Figure 2.4.

2.3.1 Analysis

After the formulation of the research question[13], the first step is to analyze it. The guiding purpose is to find appropriate theories in psychology and the social sciences[14] that are concerned with

[13] e.g. what motivates humans in general and in specific, e.g. to cooperate, act environmentally friendly.

[14] e.g. Metapsychological basics of the mind: Id (drives), Super-Ego (norms), and Ego (defense mechanims and action selection); Memory-based decision making.

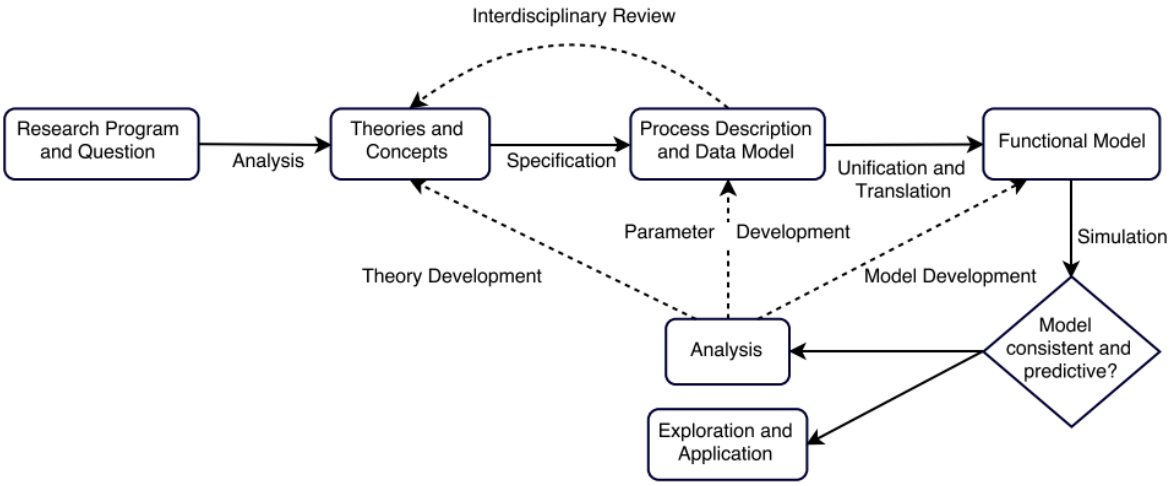

Figure 2.2: Case-Driven Simulation - Overview of Methodology.

the psychological question. Involved methods in this step are a literature research, an exploratory study (in case of weak evidences from literature), and the formulation of a so-called exemplary case (inspired by method of case studies), which follows a casuistic approach in demonstrating paradigmatic situations for the given question. This exemplification[15] of the research question supports analyzing conceptual requirements of the research question and demonstrating appropriate concepts. In such analysis, researchers from the respective field are intensively involved. All these methods support analyzing the required factors that need to be considered in answering the exemplified research question and support theory formulation. Regarding these factors as *descriptive concepts* of mental functionalities (e.g., motivations), poses sub-questions that have to be answered by theories from the respective field (e.g. the question of motivation may lead to theories about drives and norms, which lead to theories about embodiment and memories etc.). These concepts are then assessed regarding their relevance and their compliance with the followed approach (defined by conceptual criteria, see above).

2.3.2 Specification

The developed theories for the analyzed questions need to be specified and integrated into a consistent description. Therefore a process-flow description of the exemplary case is developed. The exemplary description primarily gives a behavior description using a narrative text, which must be transformed into a structured and deterministic description. The guiding purpose of this specification step is to describe the corresponding inner process-flow triggered by specified data for the exemplified situation (e.g. when reading a social media message)[16]. This is done with a so-called simulation case, which results in a process-flow and data description. That is, for a baseline scenario of the situation described in the exemplary case, (1) the required data are specified - in particular, these are the external environment, the personality (memories and personality types), and the initial state of the agents - and (2) the process-flow of decision making, triggered by theses data, is described. Since these data is supposed to determine the

[15] e.g. Describing a story of how an agent copes with its bodily needs in a social context.

[16] e.g. How do the previously identified concepts of drives, emotion etc. process the data of a given situation in the exemplary case?

model's behavior, they are termed behavioral determinants. This process is supported by the previously analyzed theories and the framework theory, which are thereby specified. In this step, typically, specification questions and assumptions arise (e.g., norms work by creating a conflict with drives).

The process-flow description can result in an implementable process model. This is the case when no experience with similar models is available or the amount of requirements is too high (e.g., to many new theories, no related question tackled previously). In such models the concepts resulting from the analytical step and their assumed processing in the mind are aimed to be represented as best as possible. This enables to analyze and identify their functional essence to develop a parsimonious functional model in a next step.

To enable empirical validation, the model must be parametrizable with empirical data. The specification of parameters supports the definition of empirical surveys. In particular, the survey questions are formulated according to the determinants specified in the simulation case[17]. This enables using surveys for direct parametrization of the model in simulations. In this case, the simulation case's standard and alternative scenarios are specified according to the survey. The usage of these simulation scenarios as test cases in software verification, then, implies model validation. Where possible, this method supports validating the model's prediction abilities. However, since surveys tend to be biased and are not able to cover all required functionalities, e.g., unconscious processes, this method is complementary, and further plausible alternative scenarios have to be specified manually in an interdisciplinary fashion.

The simulation case may be embedded in an use case description and prototype GUI (Graphical User Interface) description, which specifies the external view on the expected artifact (simulation tool). Besides user requirements, additional requirements on the model may be posed. For instance, required information in visualizations of the model's process flow or requirements for simulation experimentation may pose such requirements. This method additionally supports requirements analysis and connects the usage of the tool into the process-flow description.

2.3.3 Functional Modeling

Next, basic functions that are able to generate these processes are developed. These functions should be generic and generative. The until now separated assumptions are integrated into unified functions. This is supported by operationalizing the decision factors in the simulation case's process-flow and analyzing their commonalities[18]. Since the ultimate objective is to develop a model with functional equivalence the mind, a parsimonious model is aimed. Unnecessary details tend to impede understanding the mind's functions. However, before we are able to know what to simplify, we need to analyze the details using the previous process description. Also, analyzing the model output of a unnecessarily detailed functional model may help in finding a functional equivalent simpler model, which not only supports the explanatory power but also is easier to implement, apply, to further develop, and maintain. However, the relation to the original theories must be considered. Besides additional model assumptions, this step may include simplification assumptions.

[17] That is, which suvey questions are appropriate to represent the previously defined determinants of the external environment, personality, and internal state.

[18] e.g. drives and emotion serve valuation of data to asses how objects and actions may change the agent's state.

2.3.4 Implementation

The next step is concerned with implementing the model. First, an algorithmic and knowledge representation of the required functions and data must be found. Therefore a software design must be developed that represents the previously developed model. Of course the model functions are not identical with software functions, but the better the model is a parsimonious functional model, the clearer is the selection of required object oriented classes and interfaces. However, to implement the model additional assumptions are usually necessary, e.g., about the sequence of calling the model functions, and implementation assumptions are typically required, such as quantification of qualities and value scaling.

Where needed, as a first step an input/output blackbox model is used to approach the required software functions. That is, defining the available input and required output based on the model and analyzing the required software functions. This can be done in an iterative way: first an abstract function module with an interface description to other modules is defined, and then specified in subsequent steps. Following the implementation criteria described in Chapter 2.2.3, additional best-practices of software engineering guide the implementation step, including design patterns and agile software development.

2.3.5 Evaluation

The behavior of the resulting model is compared in simulations against the model specifications and the simulation case scenarios. This entails validating the model by comparing it against expectations from psychology and the social sciences, and (if available) against empirical data. The former validates the explanatory abilities of the model, the latter additionally validates its prediction abilities. In both cases the plausibility that the underlying theories represent a candidate explanation of the human mind's functionality is corroborated. Hence, the model's (explanatory and/or predictor) application is validated, and its representative character is corroborated.

The central test specification is given by the simulation cases specification. The model is parametrized according to the simulation case's standard and alternative scenarios and the simulation results are compared to the expectations from the simulation case. This includes testing if a change in determinants would result in the expected behavioral change. The premise of testing is the positive finalization of an interdisciplinary review concerned with evaluating knowledge translation. Further testing can be regarded as an incremental process.

The first step is concerned with testing, if the implementation is a valid representation of the model. Therefore implementation tests (e.g., unit tests on different levels of granularity) are derived from the simulation case specification. The methods used are unit tests and debugging.

The second step consists of testing, if the model is indeed able to generate the described behavior in the way specified in the simulation cases. This step consists of testing if the baseline scenario is replicable in simulation and if a change of determinants would result in the specified alternative behavior. Thus, it is tested if the parameters that are expected - due to model assumptions - to determine a given functionality, are indeed determinants. Basing this steps on determinant specifications enables to measure and test (functional) model assumptions by using data assumptions. The methods used are to observe the simulation behavior and track the causal chain of decision via visualizations. Hence, it is also tested if the model shows the correct results due to correct reasons. This step proofs that the underlying assumptions are able to be implemented in a consistent model and that they are able to generate the necessary spectrum of behavior.

If the underlying psychological theories are already established and empirically corroborated, the previous testing steps aims to guarantee their correct translation into a simulation model. In this case the model can be applied in explorations and experimentation. Otherwise a third step is necessary to compare the simulation results with real-world data. As described, the specification of alternative scenarios can be informed by empirical data. In this case, the second step implicitly considers empirical validation.

But even if empirical validation is a necessary means for real world application, it is not sufficient for model evaluation. This is not only to avoid bias from insufficient data by one or few empirical study[19], stylized facts, and over-fitting, but also to test aspects of a functional model of the mind that are empirically accessible only in a restricted form. Hence, a simulation case should consider alternative scenarios that are informed by those aspects. Such evaluation aims to involve empirical data and go beyond their limits.

If the simulation of the developed model does not generate the expected behavior, an analysis informs which artifact (theory, model, determinants) to adapt (feedback arrows in Figure 2.2 and Figure 2.4) in the next cycle of development and evaluation. This corresponds to a Lakatosian approach of doing science (see Chapter 2.1). After conduction the required iteration of the methodology's steps, the model can be further tested in unguided explorations and guided simulation experiments.

The third step would involve experimentation and exploration. Hence, in this step no test specifications are used a priori. The first step is unguided exploration, i.e., to analyze simulation data, and compare the results (correlations, clusters) with corresponding empirical data (eventually using a control group), or to check the analyzed results for their plausibility (e.g., by psychologists). This step informs the specification of simulation experiments (which can also be conducted independently). Again, the results can be compared to empirical experiments and/or interpreted by psychologists.

Since the transformation from theories to process description to model is evaluated using an interdisciplinary review process and the transformation from model to implementation by software verification, the model's validation and corroboration implies the same for the underlying theories.

2.4 Exemplary Case for (Requirement) Analysis

The abstract framework assumptions[20] of the chosen research program and a literature research guide the analysis of the research question. Here, the experience from researcher of the question's field, e.g., psychoanalysts and neuroscientists, support analyzing the required factors in tackling the research question and finding appropriate psychological concepts. Besides describing the resulting assumptions and concepts, an exemplification of them in a narrative story supports analyzing their consistence and further requirements, which only become apparent in concretization. This can be regarded as instantiating the analyzed concepts in real-world conditions. Since researcher of the research question's field are heavily involved in this step, such conretization usually include educated guesses based on their empirical experience. The less evidences a literature

[19] In this regard - to avoid following a wrong path - the simulation case and model specification should not be based on *one* empirical study.

[20] In the first step of requirement analysis only abstract framework assumptions should be considered to be consistent with the narrative form of an exemplary case (see below). Hence, assumptions about the functional-generative aspect should only be considered in the simulation case.

analysis results, the more this analysis step is required. Such casuistic way of analysis is typical in psychoanalysis. Hence, the experience of psychoanalysts in this field is harnessed for a first analysis of the research question.

Many disciplines, such as medicine and law, use typical practical cases to demonstrate their concepts. This vivid form of description can supports analysis and discussion and concretizes theoretical knowledge. In developing and evaluating a mental architecture, we use case descriptions to exemplify research questions and discuss them. Most of all the exemplary case serves as a point of departure for a common platform, bridging different disciplines. The usage of concrete narrative cases supports discussing research questions and theories between researcher from different disciplines, with different methodologies and vocabulary. Such procedure not only supports knowledge translation, but also supports conceptual analysis by analyzing research questions, supporting the explication of requirements to answer the question.

Being a first step in specifying the theoretical assumptions for the research question, an exemplary case primarily gives a behavior description. This serves reducing the complexity and supports an incremental requirements analysis. The behavior of agents are described using a narrative story. This includes describing why the agents behave this way (e.g., because of hunger), but not how the behavior is generated. For analyzing the requirements of basic human decision making and cooperation we choose an exemplary case describing how a hungry agent behave in front of another agent and a food source. Possible lines of actions are eating, sharing the food source, fighting with the other agent, or going away. To examine the requirements of basic social influence and environmentally friendly behavior we describe a social media case, where an agent perceives that a friend switched to green electricity.

The core of an exemplary case is the story, where the behavior of agents in a specific condition (environment and inner state) is described, and the corresponding context. This includes the description of their character[21]. The story, then, uses the to-be-demonstrated concepts to describe the causes for agents behavior, integrated in the storyline. To demonstrate the analyzed concepts, the behavior description should be typical for these concepts and focused on them. An analysis of the described behavior, then, pose further requirements, which are tackled in the next step of the methodology in simulation cases. Overall, the narrative part of the exemplary case should describe what happens, why it happens, but not how it happens (functionally in the mind).

Being a narrative description that embodies different theories, an exemplary case may be indeterministic, inconsistent, and may have gaps in assumptions. Hence, an exemplary case must be clarified to enable its transformation into a structured description for a deterministic model. This process includes the explication of the theories assumptions, providing a consistent and coherent description, and setting the focus and adaption of the story on the to-be-demonstrated concepts. This is supported by mentioning the assumptions behind the exemplified behavior separately and relating them to the specific part of the behavior description. If a concrete process description seems necessary at this early stage, it should also be separated from the narrative text, but still related to it. A clear demarcation about which aspects of the research questions to neglect, supports focusing on the requirements of the to-be-demonstrated concept.

Overall, an exemplary case is defined to include following information:

Research question: (Socio-)Psychological question.
Demonstrated concepts: Required assumptions, encapsulated and focused.

[21] Opposed to the term 'personality', character is a descriptive concept, describing behavior patterns.

Context: Description of agents initial internal state and character, environment. To be specified into determinants in simulation cases.
Story: Narrative description of behavior to exemplify concepts. To be structured into a process-flow in simulation cases.
Possible outcomes: Description of all lines of action and corresponding internal states, to further demonstrate the chosen concepts. To be structured in deterministic simulation cases.
Related sub-questions: Requirements as the result of story analysis. To be specified in simulation cases.
Assumptions: Explication of implicit assumptions.

Being the first step of requirements analysis and specification, the exemplary case should only provide a point of departure, but no details. This is valid for the story description, the analyzed requirements and the detail of specification. To avoid loosing focus, and reduce confusion and inconsistencies, a detailed requirements analysis and specification must be done in a structured (non-narrative) form, which is only the task of the next step, the simulation cases. Reducing the exemplary case to a basic description also enhances the clarity in interdisciplinary collaboration. Such incremental approach to requirements analysis and specification proved beneficial and was an crucial lesson learned in the process of this dissertation.

The exemplary case also provides a first analysis of the feasibility of its implementation, since the next step involves structuring it into simulation cases, which eventually will be used as test-cases. Hence, where possible the simplest behavior should be chosen that demonstrate the involved concepts and is able to exemplify the research question paradigmatically.

Such an exemplary case allows the transformation into concrete simulation scenarios, which in turn permits the specification of functions and according parameters that should generate the described behavior in simulations. Only such a case description facilitates the construction of a causal chain of description that leads to the specified behavior, a key requirement for developing a generative-functional model.

To be able to specify the requirements and model we have to transform the exemplary case into a structured and deterministic form of description, the so-called simulation case, which is an appropriate development and evaluation tool in simulation. Such simulation case serves as a shared description of the case for all involved disciplines.

2.5 Simulation Case for Specification and Evaluation

The simulation case is the second step of requirements analysis and assumption-specification. It is a structured description, resulting from a text-analysis of the narrative exemplary case.

After clarifying the exemplary case, it is transformed into a structured simulation case. In transforming the exemplary case into simulation cases, the focus is to analyze the parameters that determine behavior, i.e., *behavior determinants*[22]; in particular, how a change in these determinants

[22] 'Behavior determinants' are the conceptual term for the technical term 'parameter'. This term is preferred to emphasize that a change in these parameters are necessary and sufficient to determine the agent's behavior. However, not all parameters are determinants of behavior. Determinants are only specified on conceptual (i.e., not implementation) level. That is, determinants are the configuration of the model. The parameters specified in the simulation case are assumed to be determinants. The intensity of their determinability will be analyzed in simulations.

causes a change in the agent's behavior. Four groups of behavior determinants are specified: personality parameters, memory, environment and the agent's internal state (drives and emotions), where the first two represent an agent's personality. To structure the exemplary case, standard and alternative scenarios are distinguished. Based on the exemplary case's assumptions and clarification of its requirements (the sub-questions), the baseline scenario describes the process-flow, and the concrete determinants. Examples of determinants are objects in the environment, the agent's valuated memories thereof, parameters for the generation of drives, the distribution of neutralized intensity (which determines the priority of different function modules, e.g. planning, focusing). This analysis enable the development of a first data model.

Based on triggering by determinants, the process-flow describes in detail how these determinants are processed to result in the behavior described in the exemplary case. To support describing the process-flow, decision factors are formulated, which are influenced by behavior determinants. Since they should be able to predict the behavior - given the behavior determinants - for the subject, they are called *subjective predictors*. They should be derived from the psychological concepts from the exemplar case and, if possible, should condense them functionally. Examples are drives, norms, and emotions. The process-flow description is given by describing how these subjective predictors are changed by the parametrized determinants, and in turn how the predictors determine the agent's decision. Besides the specification of the exemplary case's conceptual assumptions, this process is supported by appropriate framework-assumptions of the research program. Nevertheless, typically specification questions arise in the specification process with corresponding specification assumptions.

A simulation case scenario, then, is specified by (1) the determinants, (2) a structured description of the triggered process-flow that bridges behavior determinants, subjective predictors, and decision, (3) the agent's final state (given by the subjective predictors), and the decided action.

A simulation case should represent all outcomes of the exemplary case that are relevant for tackling the research question and demonstrating the concepts. Alternative behaviors are described in alternative scenarios. Opposed to the baseline scenario, here it is more appropriate to conduct backward-chaining and analyze which state of the subjective predictors would lead to alternative behavior, and in turn which determinants may cause the decision factor's value. Since the simulation case's basic requirements are all included in the baseline scenario, the description of alternative scenarios focus on specifying the changes in the baseline scenario's determinants, by which the focused functionality or alternative behavior is demonstrated, and only sketch how they would change the baseline scenario's process-flow.

But alternative scenarios are not only used for alternative behaviors, but also for alternative determinations of the same behaviors. To demonstrate this concept a simulation case should involve instances of such alternative scenarios. Another usage of alternative scenarios is the demonstration of a specific functionality or phenomena (e.g., uncertainty in decision making), or to demonstrate how a specific variable (not a decision factor) influence the process-flow.

In general the method of structuring and transforming the narrative exemplary case into a structured simulation case description is inspired by use-case-based requirements analysis from software engineering. Pre-conditions of the case are represented by the determinants, post-conditions by the final state of the agent's decision factors and the selected actions. Due to focusing on enabling a deterministic description, every change in the systems behavior, resulting in alternative scenarios, must be justified and tracked causally by determinants. This includes describing how these change in determinants causes change in the system's behavior. Typically, use-cases are

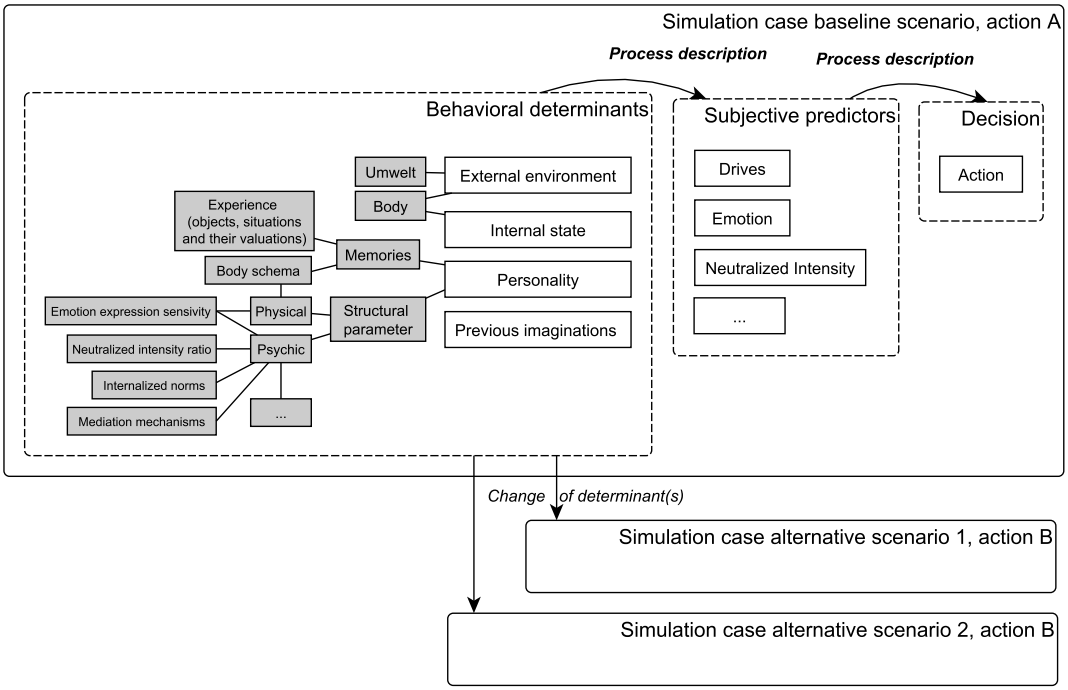

Figure 2.3: Simulation Case (SC) Structure. An SC describes how the conditions in a specific internal and external situation determine an agent's behavior. Therefor behavioral determinants representing different aspects of a situation are specified. Subsequently the SC describes how the agent's mind (i.e. the previously identified concepts) processes these specific determinants and generates a specified action. Additionally, the SC specifies how a change of these determinants is expected to lead to a change of the agent's actions.

concerned with describing the system's behavior, and based on that, requirements for the development of the system that generates this behavior are extracted. Due to the scientific nature of simulation cases additionally assumptions about how the described behavior is generated are included. All these aspects enable the usage of the simulation case as a shared representation for all involved disciplines.

An exemplary case includes the demonstration of multiple theoretical concepts. In case of a low dependence between these concepts, the exemplary case should be concretized in multiple simulation cases. This enables an iterative development, beginning with the a basic version of the simulation case, and extending it in every iteration with a focused functionality, with extending the respective determinants and focusing on the corresponding alternative scenarios. However, the decision factors must not be extended, since their principle should be covered by the basic simulation case.

Simulation cases not only support requirement analysis and concretization. They also support mapping model parameters to empirical data, which eases empirical validation by parametrizing the model accordingly. For instance, after the baseline scenario's determinants are specified, they can be used to formulate survey questions or an experimental set-up. Additionally, after empirical data is gathered, different agent types can be identified in the data and used to specify alternative scenarios for empirical validation.

2.6 Summary

The sketched methodology is an iterative process with a frequent iteration cycle of analysis, specification, modeling, implementation, software verification, and simulation, and a less frequent iteration with the analysis step (see Fig. 2.4). Technically, case-driven simulation can be regarded as test-driven modeling and implementation. Scientifically, the methodology can be regarded as bridging theory-driven and data-driven approaches. That is, we use theories as a frame to specifiy a computational model, but consider very early in specification (in the simulation case) empirical data.

After deriving a specific research question in the context of the SiMA research program, which provides framework assumptions such as meta-psychological theory, the question is analyzed by exemplifying it in a concrete situation (see "Analysis" box in Fig. 2.4). By describing how the question at stake is expressed in a specific situation in life, sub-questions arise. The methods of this step (literature review, empirical studies, and educated guesses in exemplification) result in finding required concepts that are assumed to generate the exemplary described behavior. These concepts represent psychological assumptions. To specify how a given situation in the exemplary case, represented by a specific internal state, personality, and external environment, would determine an agent's behavior the method of simulation case is used (see "Specification" box in Fig. 2.4). This step supports specifying the previously identified concepts and integrate them into a process description of how the agent's mental apparatus would work in the specified situation. This is done for a spectrum of possible behavior and situations and hence not only provides an instrument for requirement analysis and specification, but also a fine-grained test-template. To ensure a correct knowledge translation in specifying the concepts this step is reviewed by psychoanalysts and psychologists. In the next step (see "Functional Modelling" box in Fig. 2.4) information-theoretic functions are developed that are able to generate the wide range of previously described processes, in particular the wide range of beahvior by changing the determinants - as previously described in the simulation cases. After implementation, the resulting generic mental architecture is parametrized with the specific data model and tested by simulating the specified simulation case scenarios (see box "Simulation" in Fig. 2.4). Comparing the behavior of the agents and their inner workings with the simulation case specification validates the model and enable its application in simulation experiments and model exploration for the sake of answering research questions from psychology and social science (e.g. What is the impact of norms, given different states of the external environment). The whole procedure usually requires multiple cycles before a sufficient validation is reached. If the gap between specification and results is to wide, an analysis reveals which step of the methodology needs adaption (see dashed feedback arrows in Fig. 2.4). First a possible change of technical parameters are analyzed, e.g. if mapping of qualitative to quantitative values must be adapted. Second, determinant specification is reviewed with psychoanalysts, analyzing why the previously specified determinant constellation has not led to the expected behavior when processed by the specified mechanisms, which leads to the third step of analysis: checking if the model processes the determinants correctly. If all theses steps are not able to close the gap between specification and simulation, the model's underlying theories must be challenged, e.g. if inconsistencies between theories are responsible for not meeting the expectations in simulation.

In general, the described steps are conducted depending on the uncertainty of requirements (triggered by the research questions and the analyzed theories), the availability of empirical data, the need for new functionalities in the model, and the availability and consistency of theories and concepts (e.g., from previous projects in the research program). These factors determine the

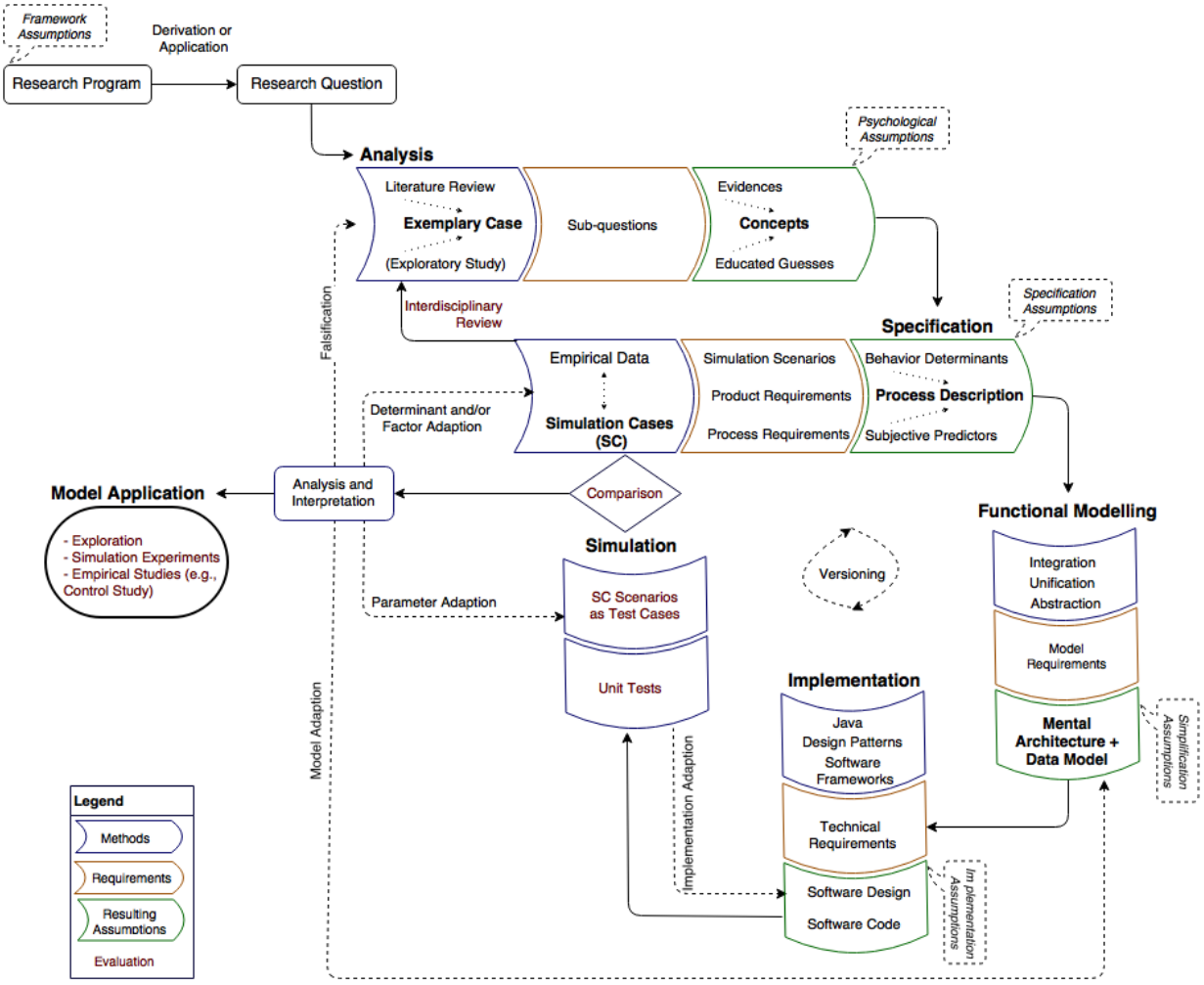

Figure 2.4: Case-driven Development and Evaluation. For a summarized description of the Figure see Chapter 2.6.

intensity of the single steps. The best case - given the required knowledge is provided - would be an omission of the analysis step and minimization of the conretization step.

The presented methodology helps to determine requirements and solutions on different levels and in a methodical way, which lead to an incremental refinement of requirements and assumptions (i.e., test specifications). Hence, requirements and assumptions of the different steps in the methodology are related to each other. Such requirements state assumptions about the features that are required to generate the described behavior. Of course, the behavior description itself represent implicit requirements, since the desired behavior is described. But we only use them to analyze the required functions that generate this behavior. Hence, we use the behavioral requirements (e.g. perceiving and focusing on a food source) to deduct – with the help of researchers from other disciplines – the functional requirements (e.g. perception for the fulfillment of desires, attention mechanisms). First the conceptual requirements are analyzed. That is, which concepts are required to generate the described behavior (e.g. motivation based on bodily needs, social norms etc.)? In most cases the process description (of the inner processes) in the exemplary case formulate the required psychological and sociological concepts. But the transformation into a technical model also may trigger requirements. Next, conceptual requirements trigger model de-

sign requirements, which the implementable agent model must fulfill (e.g. the interface between body and decision unit, environmental coupling). In developing such model, implementation requirements may feed back on conceptual and/or model requirement (see feedback possibilities in Figure 2.4). This is especially the case, since not only functional requirements, but also data requirements are analyzed. For instance, required parametrization of functions may adapt the simulation cases' defined data determinants (e.g. the incline of hunger).

3 Literature Analysis

A person who studies scientific books, aiming at the knowledge of real facts, ought to turn himself into an opponent of everything that he studies ... he should also suspect himself as he performs his critical examination of it, so that he may avoid falling into either prejudice or leniency.

Abu Ali al-Hasan Ibn al-Haytham (Latinized name: Alhazen), 965-1040 AD, "Aporias against Ptolemy".

Current psychological methodology does not encourage the formulation of comprehensive functional theories.

Allen Newell, 1927-1992, Unified Theories of Cognition.

In an interdisciplinary approach an analysis of the state of the art has to include a review of different disciplines. This serves not only to get an overview of different approaches to this thesis' questions, and supports a deeper understanding and analysis of the question, but also may serve identifying appropriate concepts to be translated and integrated into the simulation model. In both cases a description includes an analysis of the literature against the introductory mentioned criteria (see Figure 3.1 for an overview and Chapter 2.2 for their description), which will be focused in a concluding chapter.

Therefore, before digging in the specific question of how computational simulations are used to tackle the question of decision making in social context, first the general approaches of the underlying source disciplines must be analyzed. In the end, as mentioned, computer science in this thesis' approach mainly serves as a methodology to specify and examine research questions that originate from other disciplines, and hence depend on their content for model development. This is especially the case regarding the hypothesized model foundations, namely emotions.

Questions that are associated with such an approach to reviewing the state of the art are: What is the discourse in explaining social behavior? Which factors and aspects are considered in different approaches? How is computer science involved in explaining social behavior? In this regard it will be shown that the research question in this thesis' context, i.e., using mental architectures in agent-based simulations, is still an open quest with open challenges due to the lack of the introductory mentioned criteria. Nevertheless, existing models also show the relevancy of similar approaches and are not only reviewed regarding their difference to this thesis, but also regarding possible point of contacts and commonalities.

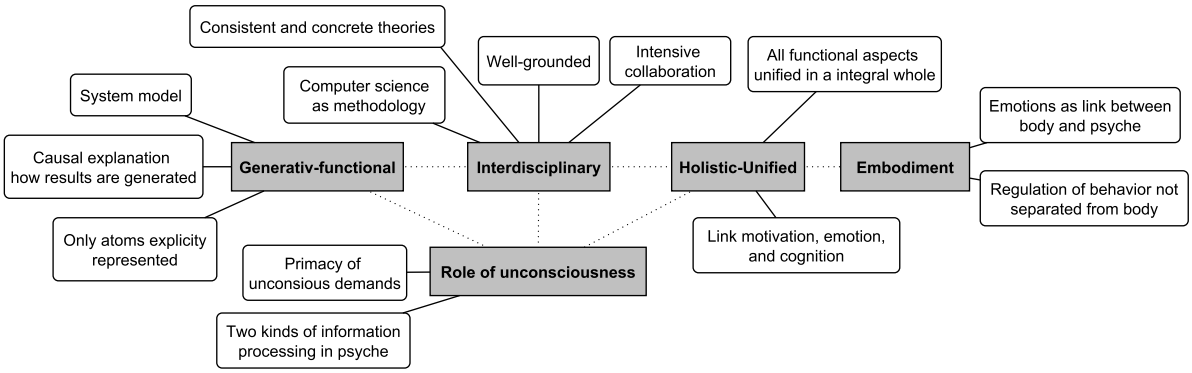

Figure 3.1: Overview of criteria for analyzed theories and models.

3.1 Explaining Behavior in Social Context: From Social Psychology to Agent-Based Simulations

Various disciplines aim to explain human social behavior. However, often these disciplines ignore each other (Sun12, p. 6), and do not follow a holistic-unified approach that integrates basic aspects of social behavior. Theories at stake are rooted in classical social sciences, foremost sociology and social-psychology, which tend between neglecting the social level (cf. radical methodological individualism: social behavior should be explained by human actions, e.g., (Via00)) and neglecting the individual level (cf. structuralism: human actions are structured by the social environment), which is reflected in different political and cultural views (e.g. Margaret Thatcher's famous quote "... there is no such thing as society. There are individual men and women." http://briandeer.com/social/thatcher-society.htm). After providing a critical overview of these streams of approaches, including a insight into the formulation of their research question, computational approaches are introduced.

With the advance of natural scientific methods and computational tools, formal approaches became prominent. An important question in this regard is, how computational methods connect and interpret classical research questions and approaches, in particular what they are able to contribute for answering old questions with new methods. In this regard it can be stated that computational approaches reflect basic questions and streams of approaches of the social sciences and humanities. This emphasizes that computer science primarily provides a new methodology (but not only as a tool), including new techniques for modeling. However, this methodology is able to enable new insights and - most importantly - is able to bridge the different streams of social sciences.

3.1.1 Theoretical Basics: Social Sciences and Psychology

Social sciences can be located between the humanities and the natural sciences, since they - dependent on the used approach - use methods from both in examining social behavior. In this context sociology is regarded commonly as the discipline that aims to figure out how human coexistence functions (SMHS09, p. 15), and generally considers bridging the explanation of living per se (alone) to living together. In an evolutionary sense creatures had to adapt on their environment, including other beings that act following the same principles as oneself. Such

adaption requires interaction; and such reciprocally adapted interactions (intentionally or not) of creatures is termed social interaction (SMHS09, p. 25). The - as a result - evolving capabilities are called social, which in the end generate a social world on top of natural conditions. This perspective represents a paradigm of sociology, which claims that such social world compete with and reduce biological determinants of human behavior (SMHS09, p. 30). Such paradigm implies the existence of (pre-human) societal factors before the human individual. Hence, it implies the dominance of social aspects, i.e. social principles enabled the evolution of humans. In other words: the adaption to the social world enabled the evolution of humans. Such sociological perspective on human thinking has primarily refer to the social world (SMHS09, p. 30). Hence, sociology conventionally uses social (not individual) structures to explain how social order arises. That is, structures that operate top-down in using entities from the social level (with concepts such as institution, group, norm, role) to the individual level. Such approach is regarded as structuralism, i.e., human (social) behavior is determined by social *structures*. Methodological individualism follows an opposite approach and argues that social phenomena has to be explained by individual actions (Via00).

Sociology has developed different theories that are used for its research methods (such as empirical methods) and for interpreting their results. These theories aim to describe and capture social reality, which is immanently autopoetic, i.e. self-organizing systems (MV87) (Luh11). In this regard social reality seems to be a chaotic process, which is influenced and determined by the interactions of various factors on different levels, in different ways (compatible, contradictory, parallel) (SMHS09, p. 83). This characteristic of social reality corresponds to the classical definition of a complex system (see Chapter 3.1.2).

To identify relevant factors, sociology developed various concepts and terms (such as group, norm etc.) which are used - in different relevance and relation to each other - in most sociological theories (SMHS09, pp. 74-83) (differently). Theories that use these concepts differently can be generally categorized regarding their point of departure and reference point in developing a model of social reality (SMHS09, p. 86). Theories may develop their model using a structural approach, i.e. reasoning about the elements that set up society. Using this perspective theories may be distinguished as macrosociologic, microsociologic, or action-theoretic. Theories may also focus on specific factors and aspects that allegedly dominate the determination of social reality. Another approach of theories reasons about the functional calculus that determines social reality. Concrete expressions of these perspectives (and combination of them) represent some of the most influential sociological theories, such as historical materialism, utilitarism, functionalism, system theory, interactionsm, and constructivism (SMHS09, pp. 84-101).

Economic answers to social questions

Economic aspects dominate some of the most influential social theories. For example, Marx's historical materialism is based on the claim that society follows the imperative of economical categories. This is also the case with utilitarism (SMHS09, p. 86), a quite general theory, which formed the concept of the utility[1]-maximizing self-interested homo oeconomicus. That is, humans decide their social interactions based on a cost-benefit analysis, and only social institutions (such as a state or the market) are able to balance such self-interested actors into a stable society. Rational[2] choice theory and game theory (Axe03) are dominant utilitarian theories. The former considers the conditions under which utility-maximization has to take place. The latter considers

[1] In such theories utility may be defined narrow (e.g. financially) or broad (e.g. some form of goal attainment).

[2] The term rational is used regarding utility-assessments. In this sense, only decisions that maximize the benefits are regarded as rational.

an actor's limited consideration of other actors' possible decision. Modern rational choice theories, such as the RREEMM (Resourceful-Restricted-Evaluating-Expecting-Maximizing-Man) model (Lin85) (Sch09) - opposed to conventional approaches - also start to consider restricted action opportunities and resources.

One reason for the current dominance of rational-choice approaches is their allowance of formalization and deduction, but critics emphasize that their unrealistic assumptions constrain their real world usage to a degree that such theories are often not usable as a prediction model (Axe03). These quite abstract principles of the rational-choice paradigm can be regarded as preventing its falsification. However, various arguments are brought up against conventional rational-choice theories, which can be subsumed by Herbert Simon's theory of bounded rationality (Sim56). The general point here is the question on which basis a person decides its cost-benefit calculation. An objective consideration of all necessary information is not realistic and psychologically not plausible due to humans' (cognitive) restrictions. Instead of aiming for utility-maximization, theories of bounded rationality aim for utility-satisfaction. That is, the decision making process is aborted, when a (under the current conditions) satisfying alternative is found.

Towards behaviorally realistic models of decision making

Herbert Simon's theory of bounded rationality (Sim56) implicitly established the quest after unconscious mechanisms of decision making in various fields[3]: Since then his theories are elaborated in various disciplines, from economics to cognitive sciences, and various researcher have demonstrated that humans are not maximizers, but satisficer (a concept coined in (Sim56), joining the terms 'satisfy' and 'suffice'). Impressive examples of how most of our decision are determined unconsciously is given by Bargh and Chantrand in their seminal work (BC99). By showing the limited abilities of complete conscious processing they challenge this dominant view, and identify three major mechanism of 'automatic self-regulation' (BC99): an automatic effect of perception on action, automatic goal pursuit, and a continual automatic evaluation of one's experience. In their experiments Bargh and Chartrand also demonstrate how these automatic mechanisms, which proved beneficial in evolution, are gates for bias and manipulation. This downside of our decision making process is vividly demonstrated by Ariely (Ari00)(Ari09), which shows how predictable 'irrational' (in the sense that a subjectively chosen decision may be objectively harmful) these mechanism may make our decision.

Unconscious and automatized decision making

The processes behind bounded rational choice may be distinguished according to their processing-principles, often termed conscious and unconscious thinking. (Kah11) calls them System 1 and 2. Whereas System 1 operates automatically and quickly, with decreased effort and no voluntary control, System 2 jumps in for effortful mental activities that demand focussed thinking. System 1 works under the conditions of cognitive ease (Kah11, p. 60): repeated experience, clear display, primed idea, and good mood. If these conditions are not met (e.g., when the person is in a bad mood or has no experience with the situation etc.) System 2 comes to the fore. System 1 operates via associative thinking, which is triggered by perception and activates similar memories. One term to describe such processing is 'intuition', which can be regarded as a form of recognition. "The situation has provided a cue; this cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition" (Kah11, p. 11). However, Kahneman distinguishes between heuristics as a rule of thumb used in decision making under uncertainty, and accurate intuitions of expert,

[3]Herbert Simon and Daniel Kahneman are Nobel laureates in behavioral economics.

which become automatized by practice. Both are done in System 1, but on different (data) basis. Overall, corresponding information processing in the mind is using our experience to have all relevant data for the future, under consideration of the constraints human thinking (Kah11, p. 51). The relation between the two systems can be regarded as System 1 generating suggestions ("... impressions, intuitions, intentions, and feelings ... " (Kah11, p. 24)) for System 2, which turns them into beliefs and voluntary actions. However, System 2 reflect on System 1's suggestion only when necessary, i.e., when the conditions of cognitive ease are not met. Hence, System 2 only operates with other principles, but still only with data of System 1 (Kah11, p. 85). However, concepts such as intelligence are not solely anchored in System 2, since it is also "the ability to find relevant material in memory and to deploy attention when needed" (Kah11, p. 46), which is done in System 1. The usage of heuristics enable to deal with complex questions by finding related, more easier questions, in case of not finding a satisfactory answer quickly.

To demonstrate the principle of these two systems and the dominance of System 1, Kahneman and Tversky used experiments that demonstrate the relation of perception and intuition (TK74). In particular, how default decision making is processed by associative activation of memories through similarity. This principle of heuristics results in often unreasonable decisions - in terms of statistics. However, as Kahneman showed, since System 1's processing cannot be described by statistical terms, and most of our actions are decided by System 1, such statistical 'irregularities' are common in humans. A demonstrative example goes as follows (Kah11, p. 156 ff.): Linda studied philosophy ten years ago and was very active in anti-nuclear demonstrations and concerned with social justice. When asking probands what they think Linda is more likely doing today, bank teller or feminist bank teller, most of them choose the latter.

Heuristics as rules of thumb can be categorized as anchoring, availability, and representativeness (TK74) (TS09, pp. 24 ff.). Anchors serve as known information that guide decision making. In the 'availability heuristic', salient information guides decision making. Representativeness is concerned with aligning new information to the most similar prototype of a category.

The evolution of heuristic can be explained by the higher importance of risk avoiding compared to gaining something.

In decision making less is (often) more

Some researchers in principle endorse the relevance of heuristics, although they assign them different key features and conditions when they are used. For example, Gigerenzer defines heuristics as "fast and frugal" operations (GG96), which Kahneman contests, since he argues that frugality is not a defining element of heuristics, given the parallel processing capabilities of large amount of information in the brain (Kah13, p. 404). According to Gigerenzer, frugality is a beneficial key element of heuristics because it avoids over-fitting (Gig08, p. 84ff.). That is, in most cases less information is more adaptive to the current situation, if this lesser amount of information is more relevant. This avoids adapting (or fitting) the decision on a big amount of tendentially irrelevant information. One perspective to consider this is that the order of activation of information should determine its prioritization, instead of regarding all data as equivalent and try to optimize the decision. This view again contests the conventional approach of humans as maximizing their expected utility, which can be regarded as a balance-sheet-method weighting all alternatives with their 'pros and cons'. Instead an approach that 'satisficies' humans expected utility (cf. (Sim56), see above), which can be termed as 'Take the Best' heuristic, is psychologically more plausible, evolutionary more probable, and more adaptive. Gigerenzer (Gig08) demonstrates why this is the case: Fitting your decision on all available information from your experience would not be adaptive to the current situation; this can be regarded as a hindsight fallacy (opposed

to hindsight bias). Usually, the experiences activated first, including the previously activated information, usually is the more adaptive one. This implicitly emphasizes the key role of perception and environment. In this regard, perception activates memories that are adapted on the current environment. And the suitability of a heuristic can be measured by its adaptivity on the environment. An suitable analogy is a comparison of human behavior with a scissor, whose blades are the structure of environment and the person's mental processes (Sim90). Another vivid analogy, which also demonstrate that the complexity lies in the environment, is that of ants (Sim96): The behavior of an ant finding its path through a beach full of twists and turns seems to be complex, but the complexity lies in the environment, not in the ant. To understand her behavior one has to consider the structure of environment and the structure of her information processing. The environment, then, is a key condition to determine if and which heuristics are used. In this regard heuristics are defined as decision processes in an uncertain environment ("In an uncertain environment, good intuition must ignore information" (Gig08, p. 85)).

Usually the concept of heuristics is used in a descriptive form. That is, researcher show the dominance of heuristic thinking, but seldomly relate them to psychological concepts. And if, they only do this abstractly, e.g. defining recognition, imitation, emotion as evolved capabilities that are used by heuristics (Gig08, p. 58). However, this gap of defining the building blocks of heuristics, e.g. how (heuristic) decision theory links with theory of emotion (Gig13, p. 50), are accepted challenges in the field.

3.1.2 A Computational Method for Social Sciences

Since the rise of formal models (with the establishment of computers) to tackle questions in the social sciences, different approaches were developed. Prominent examples are system dynamics (e.g., Club of Rome's World Model (Mea72)) and statistical models. They represent analytical models, since they try to represent social dynamics, i.e., macro-phenomena (e.g., crowd behavior), in a top-down manner. Often such models are described as "God's-Eye-View" approach (BT11) to model a global system (the society). The classical idea that a system's behavior can be deduced a priori by solving equations turned out not feasible, since they are computationally irreducible (BT11). This is generally the case with complex systems, which cannot be described on a system level (since not enough global information about such systems can be obtained), but only by describing their single components. This distinction can be regarded as reflecting the dichotomy between a sociological and psychological approach of describing social behavior.

Complexity

Until the late twentieth century scientific theories have seldomly taken Aristoteles' catchphrase 'The whole is more than the sum of its part' seriously. Instead the reductionist approach in science assumed that the whole can be understood completely, if you understand its parts and the nature of their sum (Mit09, p. IX). One of the first who provided a structured account against such an approach, using complexity theory as a counter approach, was Herbert Simon (Sim65). He identified high hierarchy and near-decomposibility as a features of complex systems, i.e., systems that are composed of subsystems, where each subsystem is highly intertwined (cf. the etymology of complex as entwined).

Approaches to measure the degree of complexity in systems are quite vague, e.g., by measuring how difficult a system is to describe, by how difficult it is to be built, or by its degree of organization (Llo01)(Mit09, 95). However, the difficulty of grasping complexity is immanent to the

concept. The most common definition includes non-linearity as key feature of complex systems. That is, the effect on a dependent variable is not proportional to the sum of effects by independent variables. This implicitly means that an a priori static defined relation between the variables of the system does not hold linearly in the operating system: The relationship between the system's variables is dynamic, i.e. the effect-relation between the variables changes on run-time of the system. Complex systems generate emergent phenomena, which have properties that are decoupled from the properties of the system's parts (Bon02). Such systems are also regarded as chaotic, with a seemingly random behavior emerging from a deterministic system, and the impossibility of predictions due to sensitive dependence on initial conditions (Mit09, 38). So, in complex systems an analytical approach that tries to understand the system top-down, i.e., from a (global) system description, is not feasible, since a description of the system derived from its behavior is not possible. That is, complex systems are not mathematically describable on the (global) system's level.

The meta-field of complexity sciences aims to find general principles that hold for all complex systems and helps to explain and predict them. Central insights of complexity science can be described as follows. Assumed reasons for complexity, such as feedback loops (Bon02) and the frequent interaction between a system's individual entities, imply the necessity of a bottom-up approach to represent complex systems. Due to the insufficiencies of an analytical approach, the requirement of a computational approach is to generate and explore a system's behavior. In this regard, since the behavior of complex systems is not predictable (a priori), it must be generated in simulations. That is, we have to develop and run computer models and observe their behavior to predict and understand the global properties of such systems (BT11). The most common approach for that is agent-based modeling, which aim to link the behavior of a system's individual entities with the global system-behavior. This is regarded as bridging the micro-macro link in ABSS.

Agent-Based Modeling

We saw that social systems are claimed to be only describable in a bottom-up way through their components' (interaction) due to their complexity. After increasing availability of computational power, various forms of bottom-up approaches, which represent the system by modeling its atomic behavioral entities, were developed and simulated. After developing micro-simulations ((Gil07, p. 57ff.)) and cellular automata ((Gil07, p. 130ff.)), agent-based models (ABM) in agent-based social simulations (ABSS) were established as the currently most used bottom-up modeling approach in social simulations. In ABSS the modeled entities are called agents, which are autonomous and heterogeneous (i.e., they are parameterized individually) individuals. The system's behavior emerges by the frequent interactions of these agents. Such an approach, which explains social phenomena by 'growing' it through agent interactions, is termed generative (Eps99). The generativist's question, then, is "How could the decentralized local interactions of heterogeneous agents generate the given regulatory?" (Eps99, p.41). Using the ABM method, this question is answered by "Situate an initial population of autonomous heterogeneous agents in a relevant spatial environment; allow them to interact according to simple local rules and thereby generate - or 'grow' - the macroscopic regulatory from the bottom up." (Eps99, p.42). The simulations then show if the hypothesized micro-specifications suffice to generate the targeted phenomenon.

ABSS proved to be suitable to represent complex social systems, since the key approach to represent them is to consider the interaction of their components, or to paraphrase Aristotle: "The whole is the sum of its parts *and their interactions*" (Bon02). Additionally, ABSS is a more natural way of a system's representation, and therefore supports realistic parametrization,

experimentation, and mapping from the real to the simulated world. This aspect also emphasizes why ABSS is used as an interdisciplinary approach. But the appropriateness of ABSS in social sciences is also argued on mathematical ground: (BT11) argues that ABSS, as a blend between classical and constructive mathematics, is the more appropriate form of mathematics for social sciences.

ABSS is mainly involved in demonstrating macro-level dynamics with simulating the micro-level and its interactions, and therefore bridging the gap between the two levels. In most cases such simulations are involved with formation dynamics, e.g. distribution of wealth, price development, epidemiological spreading, segregation, norm development. (AT06) lists: "(i) Emergence and Collective behaviour, (ii) Evolution, (iii) Learning, (iv) Norms, (v) Markets, (vi) Institutional Design, and (vii) (Social) Networks as examples of application areas." Paradigmatic models of bottom-up social simulations are Schelling's segregation simulation (Sch69), Axelrod's Iterated Prisoner Dilemma (SAS94), and Epstein's simulation of the Anazi culture (Eps99), which will be described next.

One of the first models that generated an emergent phenomena using a bottom-up method was Schelling's segregation model (Sch69). It argues that segregation occurs even though individuals do not act in a coordinated fashion to bring about the segregated outcome. Intuitively, one would expect that segregation only occurs if a person-type aims to live in neighborhood dominated by his type. But the model shows that even a 50% acceptance rate of other typed persons in a neighborhood-lattice would - eventually - generate separation. Hence, you cannot explain the macro-level phenomenon of segregation with intolerance (i.e., a preference to one owns person-type) in the micro-level. However, given an odd-numbered neighborhood-lattice, Schelling considers the neighborhood-evaluating person as an observer that does not include herself in calculating the type-rate in a neighborhood. Nevertheless, the model provided insigths by demonstrating that total segregation emerges even with a mild preference of one owns type.

Although ABM is mostly used to explain macro-level phenomena, it is also used to tackle organizational, group and interpersonal phenomena. A classical example of the latter are game-theoretic simulations of the Prisoner's Dilemma, which was - in its today used form - formulated in (Pou92) and is a typical example of game-theoretic thinking in the cold war. Stated as an interpersonal conflict it is used to explore cooperation in general. The Dilemma is described as two prisoners, who are interrogated separately. Each prisoner has the opportunity to testify against the other and get a decreased sentence. However, if the two prisoners would not testify (i.e., they cooperate), the prosecutor would not be able to convict any of them. But since a prisoner does not know if the other agent also cooperates, the dominant strategy is to testify against him. Hence, the Prisoner's Dilemma is a paradigmatic case, where from a global perspective cooperation is the best choice for all, but from a local (i.e., personal) perspective, defection is the best (i.e., rational) choice. It can be regarded as describing the conflict between individual and collective rationality (Sch09, p. 39). For many people, it is not surprising that humans often do not decide objectively rational. Additionally, game-theoretic examples usually have the same flaws as conventional rational-choice-models (see Chapter 3.1.1), i.e., they do not represent humans' decision making realistically. However, from an objective point of view the question remains why humans then cooperate at all. To explore how - opposed to the findings of the Prisoner's Dilemma - cooperation between selfish individuals evolved in nature, without being institutionally forced, Axelrod (Axe06) developed a simulation tournament with an iterated Prisoner's Dilemma scenario, where multiple runs of the game are played and the prisoners remember the other's decision in the next run. Surprisingly the most simple tit-for-tat algorithm won, which decides per default to cooperate on the first iteration of the game, and after that, imitates the opponent's move. This

outcome emphasizes, as typically, that ABSS proves appropriate to simulate frequent interactions between persons, in particular by giving counter-intuitive insights.

Overall, one key challenge in ABSS is how to develop the required micro-specifications. Since conventional agent-based models are primarily interested in the micro-macro link, sophisticated micro-specifications, i.e., a psychologically plausible decision model, are conventionally neglected. Additionally, given the immanent counter-intuitive feature of emergence, the systems' behavior cannot serve as a point of departure for deducing micro-specifications. These constraints are the reason why ABSS researchers talk about the "art of choosing appropriate micro-specifications" (Bon02). Only the usage of psychologically plausible models can tackle this challenge and show how social dynamics are generated with dynamics on the psychological level.

3.1.3 Towards a Cognitive Turn in Social Simulations

The discussion between sociological, economical, and psychological approaches is reflected in the discourse between social simulations (ABSS) and cognitive simulations (e.g. cognitive architectures). Similarly to sociology, conventional ABSS argue that macro-level phenomena are better explored using abstract micro-specifications. A typical argument against cognitive decision-making models include that sometimes an agent's properties at the psychological level are assumed to be constant. Another argument is that often the processes that give rise to the agent's action (and their alternatives) are not relevant (Gil06, p. 430ff.). That is, in such cases it does not matter why the agent decides the action, relevant is only what he does in the given context (e.g. in Schelling's segregation model (see above)).

Recently, different researchers in ABSS and cognitive science (e.g., (Sun06b), (Cas98a), (JJ03) etc.) are arguing against the usage of simplified, psychologically implausible decision models in ABSS. Typical arguments are that many macro-level phenomena are caused by the dynamics between the micro and macro level, which is not captured by (too) simplified models (JJ03). Another key argument is that replicating macro-level phenomena in conventional ABSS does not imply that the relevant micro-specifications are considered in the model. In extreme cases, the 'right' phenomena are replicated due to 'wrong' assumptions. Also, simplified abstract micro-specifications are difficult to compare against established psychological models or against empirical observations (Sun06b, p. 17). This limits the usage of conventional ABSS in real world applications (e.g., for policy testing). ABSS supports understanding the dynamics in social simulations and identifying emergent, i.e., unpredictable, phenomena. Hence, ABSS is helpful as a starting point to identify turbulent states in complex systems, but is limited in addressing them. For instance, to be able to compare the model's parameter with empirical data and in turn inform policies of how to tackle the phenomenon with psychologically-grounded variables, behaviorally realistic models are necessary (JJ03). In the end a tradeoff between abstraction and specification of the decision making process is necessary, since on the one hand a model implies abstraction, but on the other hand we cannot know the causes of a phenomenon a priori, but we can tackle the structure that generate these causes, which is the human mind. This is particularly the case with counter-intuitive emergent phenomena. One possibility for a trade-off is to use basic building blocks of cognitive models as foundations and, dependent on the concrete research question at hand and insights about emergent behavior from prototypical simplified ABSS, to decide on their detail of specification and extension (SJD16).

The traditional necessity to have an explicit model for social phenomena decreased significantly with todays computational means and information models (SJD16). Most importantly, today's

computational power enable large-scale social simulations with sophisticated models. And they enable the integration of social and psychological aspects in simulated decision-making, which support deeper explanations and better forecasting. In particular, it provides deeper process explanations which contribute to a more adaptive management of less predictable (complex) developments. As general in science, top-down or bottom-up approaches in integrating social and psychological levels in simulations are possible. A top-down approach would start with a simplified social simulation, considering abstract cognitive building blocks, and subsequently informs the necessity and detail of specifications in the psychological layer using simulations (SJD16). Additionally, docking from the social level to the psychological level can be a difficult task, in particular due to the counter-intuitive aspect of emergence. The bottom-up approach would start with representing a psychologically plausible representation of human decision-making, since it is the source of all phenomena, and identifies the (unexpected) causes of the social phenomena (SJD16). Here, simplification would be done after exploring the model in simulations. A bottom-up approach – starting with assumed building blocks - would also resemble a generative approach, which is one of the core-theme in ABSS, but such a detailed approach may lack feasibility.

Overall, ABSS is a key methodology and scientific framework to tackle complex systems and to explore emergent phenomena, in particular to close the micro-macro gap. But simplified models are only the first step, since they do not harness the full possibilities of the agent-based paradigm, and impede the causal comprehension of the micro-macro gap. Again, similar to the lack of communication and integration between sociology and psychology, social and cognitive simulations should not exclude but complete each other. For instance, in conventional cognitive models the impact of social structures (e.g., institutions) is not considered sufficiently (see Chapter 3.2). This would extend the core explanatory power of ABSS into a form of cognitive emergence (CE) (Cas98a).

3.1.4 Excursus: (Cognitive) Emergence

The concept of Cognitive Emergence (CE) is used to describe the emergence of social dynamics out of social interaction between cognitive agents (Cas98b). In this context cognitive agents are agents that have internal representations about their beliefs and goals which can be manipulated and generated internally. The concept of CE is used to describe "when agents become aware, through a given 'conceptualization', of a certain 'objective' pre-cognitive (unknown and non-deliberated) phenomenon that is influencing their results and outcomes, and then, indirectly, their actions" (Cas98b). In other words, the objective macro level expression is explicitly represented in the persons mind with subjective opinions and theories about it that consequently lead to corresponding actions. This transition from objective to subjective, from implicit to explicit, from unaware to aware, which is an intrinsic aspect of cognitive perception in social interaction, characterizes CE. Considering such social interactions that are objective relations between cognitive persons, i.e. "they hold independently of the awareness of the person" (Cas98b), one can observe that cognitive emergence may strongly change the social situation. This manifests in feedback, respectively the influence of the subjective interpretation of an objective structure on a cognitive agents feedback, in the context of CE defined as immergence (Cas98b).

Since the concept of emergence is used in different ways, it is necessary to relate the usage of CE to it. As stated in (Eps99) emergence is often used with vagueness and sometimes with mysticism. There is even an anti-scientific flavor in such an usage, if in-explicability for emergent phenomena is implied. When de-mystified, such approaches can mean nothing more than that

the emergent phenomenon may *presently* be not explainable (Eps99). The goal in science should be to fill the gaps, not to preserve an explanation gap between different levels of description (e.g., micro and macro level). This must be considered when using the concept of emergence. In ABSS the concept of emergence is usually used to describe the observation of a structure or feature on the macro level, which is a result of interactions on the micro level. ABSS is a flexible possibility to analyze and explore emergent phenomena by changing the parameters of the concerned entities (assumptions about the micro level), respectively of the simulation. Hence, agent-based simulations are especially helpful for the simulation of social structures where the interaction of the agents is the main topic of concern.

A key aspect of emergence is the transition of sight and view (Cas98b). This implies the revelation of a previously hidden phenomenon with the change of observation level and hence description level. This notion complies with the characteristic of CE regarding the transition from objective to subjective, from implicit to explicit, from unaware to aware. So, regarding this key-characteristic of emergence, the concept of CE clearly relates appropriately to it. For another key-characteristic of emergence this is not unambiguously the case. In most definitions of emergence the term is used for the macro-level, i.e., the emergent phenomenon is observed on the macro-level. In CE this is not so obvious. On the one hand we can observe that the transition from objective to subjective happens with cognition and hence on the micro-level. On the other hand the emergent phenomenon is observed by its impact on the macro level, whose primary scope is the relation between the interacting cognitive agents. In this regard the point of observation has to be defined. Obviously without observation the distinction of different description-levels of a system or structure, and hence the detection of emergence, is not possible. In this regard in the field of ABSS a distinction is made between strong and weak emergence (DMP07). If the agents are able to observe the emergent phenomena in which they are involved one refers to strong emergence. If the observer is external to the system, the phenomenon is described by weak emergence.

The need for an external observer to detect emergence directly (and not via immergence) can be avoided by using of an alternative definition of emergence, namely intrinsic emergence. This definition suggests to characterize emergence as "an autonomous increase in the system's computational capabilities" (Ber04) (DMP07). Hence, this definition is not dependent on the emergence of structural patterns and on the detection by an external observer. Such a definition is supposed to be more objective and generic. Intrinsic emergence may be seen as behavioral emergence, because of its discontinuity in its behavior. Thus, intrinsic emergence, in opposite to structural emergence does not rely on the complexity of structure. With such characterization of cognitive emergence and relation of it to a compliant definition of emergence it is obvious that it is not sufficient to use a broad concept such as emergence without relating and embedding the usage of it to a compliant definition. In this regard one can state that CE is not the form of emergence that is widely used in ABSS, namely the emergence of a structural phenomenon on the macro-level, detected by an external observer. It is rather a special form of emergence that still complies with the essence of the broad concept of emergence.

3.2 Cognitive Models for Social Simulations

Current limits of ABSS show that deeper explanations and realistic predictions require models of human decision making. Some approaches extend their simplified model, others use given models. When aiming for a cognitive turn in social simulations, cognitive architectures are an alternative. Originally not conceptualized for social simulations, they are adapted for their usage

to include social aspects. Before looking at how cognitive and psychological models are used in social simulations, cognitive architectures are introduced and evaluated against the criteria of this thesis. Particularly important criteria in this context are the consideration of the human mind's fundamental principals and their integration with high-level cognition into an unified model. Another important line of review is how cognitive architectures consider social aspects.

3.2.1 A Computational Method for Cognitive Science

A central principle of Cognitive Science is the integration of disciplines, and hence different levels of explanation, to explain the human mind's functioning. This implies the requirement of a framework, which can be regarded as *the* job of Cognitive Science (Ber14, p. xxvii). In Cognitive Science the mind is usually regarded as an information-processing system (Ber14, p. 3)[4]. Hence, a logical consequence (given the Zeitgeist) is to use concepts and methods of the field that are concerned with information processing in a general sense: computer science. Hence, it is not surprising that some of the founding fathers of Cognitive Science, such as Allen Newell, were computer scientists. So, if the human cognitive system can be regarded as an abstract computer, similar to specifying a computer *architecture*, the goal of Cognitive Scientists should be to develop cognitive *architectures* (BN71)(And07, p. 5), i.e., a specification of its structure and how it achieves the function of cognition.

Thus, cognitive architectures could serve as a general framework for Cognitive Science and provide the means for an unified theory of cognition. However, as often, means determine theories. This is also shown in the (original) connection between Artificial Intelligence and Cognitive Science. In fact, the point of departure was to develop intelligent machines that are capable of solving complex problems. This drove the focus on reasoning machines, which are reflected in typical AI application fields of chess and algebra. The Cognitive Science perspective on this approach led to a generalization of their aim, e.g. to develop a general-problem-solver (GPS). Such approach aims "... to construct computer programs that can solve problems requiring intelligence and adaptation, and to discover which varieties of these programs can be matched to data on human problem solving." (NSS59, p. 1). Typically for this approach (and for later cognitive architectures, such as ACT-R, see below), algebraic problems are used as prototypical problem statements. The usage of such approach was argued also for psychology "... to construct complete processing models rather than the partial ones we now do." (New73, p. 18).

If the human mind (as a GPS) is able to solve complex problems, such as chess or algebra, AI and Cognitive Science have the same objective, namely to find the structures behind these capabilities, i.e., the architecture of cognition. The initial focus on complex problems (given the influence and mindset of computer scientists) was probably a reason why cognition is traditionally viewed as higher-level problem-solving. This is reflected in the common definition of a cognitive architecture as a blueprint for intelligent systems (e.g., (DOP08)(LLR09)). The characteristic feature of cognitive architectures (with commonalities to the brain) is to model a generic system, i.e., to find the structural properties on a system's level. These are exactly the building blocks "... constant over time and across different application domains ..." (LLR09, p. 1).

On a system's level a functional analysis of the required components of a cognitive architecture is typically done top-down (in computer science). Such approach is best done on a symbolic

[4]It has to be mentioned that information or information processing are often used generically and a common definition does not exist in Cognitive Science.

level. And using a functional perspective, only the abstract functions are relevant, not their implementation. In this regard a symbolic approach is sufficient, without having to know how the functional system is implemented, as long it is implementable physically.

This principle is rooted in the physical symbol system hypothesis (New80, p. 170): "The necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system", with defining general intelligent action as "... the same scope of intelligence seen in human action: that in real situations behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some physical limits." (New80, p. 170). Physical symbol systems can be characterized with four basic ideas (Ber14, p. 143): "... (1) Symbols are physical patterns. (2) These symbols can be combined to form complex structures. (3) The physical symbol system contains processes for manipulating complex symbol structures. (4) The processes for generating and transforming complex symbol structures can themselves be represented by symbols and symbol structures within the system". In a nutshell the most important process in such systems that give rise to intelligent action is claimed to be querying a search-space using a heuristic approach.

This definition again emphasize the computational roots of cognitive architectures. The most prominent argument against such a 'strong AI' approach that aims to develop artificial minds (opposed to weak AI that tackles narrow problems) is depicted in 'The Chinese Room Argument' (Sea80), which aims to demonstrate that no symbol manipulating system is able to justify the claim of replicating human intelligent action.

Another argument against such a so-called cognitivist approach is that physical constraints do have impact on the algorithmic symbolic level, e.g. time constraints (the physical implementation is a determiner of the computation speed). This leads to following a bottom-up approach that aims to generate the capabilities of the mind using artificial neural networks (ANN). Although such ANN are only biologically plausible from an abstract perspective, they aim to bridge the gap between the implementation and the algorithmic level (RMG86). However, they are not appropriate to explain the algorithmic and functional level of cognitive systems. But according to followers of such so-called connectionist (or emergent) approach, not the explicit function of the mind and symbols should be modeled, but their substrate: neurons and their connections. The function of mind would then emerge from such implicit models. However, when using ANN to replicate the human mind, limited insights could be drawn, like any model that fails to abstract from irrelevant aspects. Approaches with such claims are called emergent, opposed to cognitivist approaches, which claim that emergent approaches are only implementation models for physical symbol systems (Ber14, p. 270) .

Thus, the question is which level of organization to choose in order to represent the human mind? This leads us to another definition of a cognitive architecture: "A *cognitive architecture* is a specification of the structure of the brain at a level of abstraction that explains how it achieves the function of the mind" (And07, p. 7). But what is the right level of abstraction in representing the mind? Probably there is no *one* answer to that question, since the level of abstraction cannot be fixed for such a general question, but must be adapted to specific instances of that question, e.g., for the different functions and tasks of the mind.

Overall, emergent systems are not able to replace (symbolic) cognitive architectures, and vice versa. Both have explanatory power in different domains that should not compete, but complete each other. Both are necessary for a model of human mind, but none alone is sufficient. ANN proved appropriate for sensor-based processing, such as pattern recognition in perception, and cognitive architectures proved appropriate for reasoning. This insight led to hybrid cognitive

systems that considers the right level of abstraction for different functionalities of the human cognitive system, i.e. follows a multi-level approach.

An extend definition that considers the requirement of different levels (or domains) defines a cognitive architecture as: "... a domain-generic computational cognitive model that captures essential structures and processes of the individual mind for the purpose of a broad (multiple domain) analysis of cognition and behaviour" (Sun06a, p. 33).

3.2.2 ACT-R: A Production-system-based Cognitive Architecture

ACT-R (Adaptive Control of Thought-Rational) is a typical example for a cognitivist architecture. It is currently one of the most referred cognitive architectures and claims to represent a consensus in cognitive science (And07, p. 55). The cognitive architecture is described as "best grounded in the experimental research literature" (Mor03, p. 24). This is reflected in the history of ACT-R. It draws strongly on the idea of production systems and means-end-analysis, associative memory, bayesian adaptation, perceptual-motor coupling, modular architectures, and the difference between sub-symbolic and symbolic representation (And07).

The ACT-R main components consist of modules, buffers, and pattern matcher. Its central paradigm is a modular structure and the integration of these modules. The central question for ACT-R then is how "... cognition emerges through the interaction of a number of independent modules ..." (And07, p. 19). ACT-R distinguishes perceptual modules (visual and aural), response modules (manual and vocal), and central processing modules (imaginal, declarative, procedural, goal). The functions of these modules can be exemplified with a typically used example in ACT-R: algebra, e.g., solving $3x - 7 = 5$. The imaginal module would hold a current mental representation of an intermediate equation, e.g., $3x = 12$. The declarative module would fetch relevant factual knowledge, such as $7 + 5 = 12$. The goal module would keep track of the goal state, i.e., how to solve the problem, e.g. the next step to do. Hence, this module provides the ability for means-end analysis. The procedural module would provide general rules for solving equations (And07, p. 53). The modules communicate by writing into their associated buffers, with each buffer having the capability to hold one so-called chunk (a measure of information amount). This represents one of various cognitive constraints that are purposely built in the architecture. The integration of modules is achieved by the procedural module, which queries the production rule that fits best to the pattern of combined chunks in the different buffers, and applies the rule to manipulate the chunks in the buffers[5]. Here again cognitive constraints are built in, with the limitation that only one production rule is allowed to be executed (opposed to other production systems) and a fixed duration of 50ms for execution (which is assumedly the time the corresponding brain region needs). Overall, the procedural module is designed as the central bottleneck in information processing. Besides conceptual reasons, these limitations have functional purposes. For instance, constraining one production rule to fire avoids contradictory manipulations of multiple rules on buffers (And07, p. 61).

ACT-R follows a strict modular approach, which "... seems the only way to achieve the multi-purpose functionality that humans need to meet the demands of their world given the structure of their brains" (And07, p. 86). This reason originates in a philosophical (and evolutionary) deduction that the human mind must follow a modular organization to be able to display its

[5] A production rule in this context could be described as: "If the goal is to solve an equation, and the equation is of the form 'expression-number1=number2,' then write 'expression=number2+number1"(And07, p. 21)

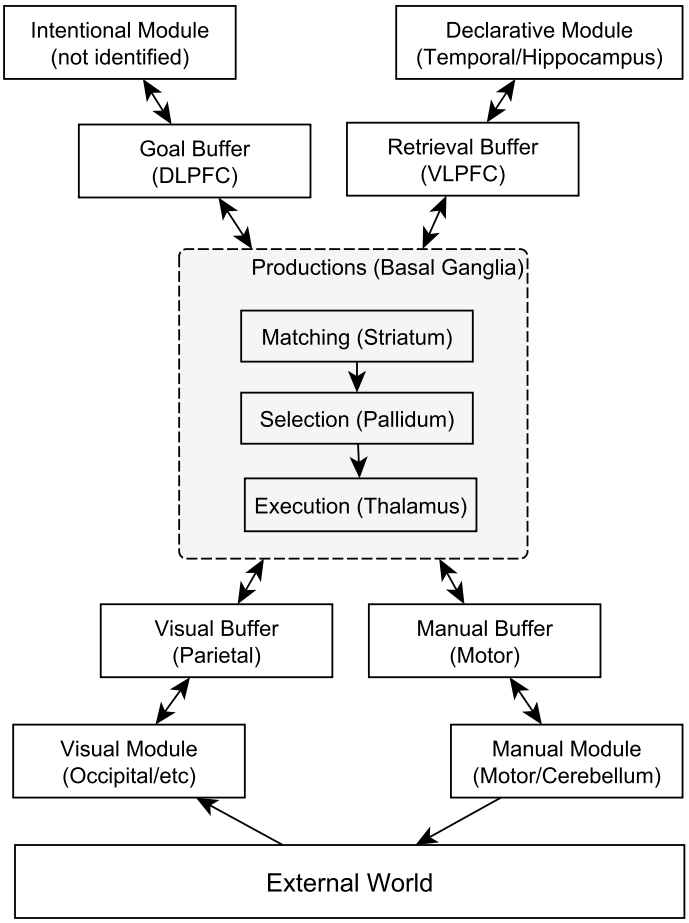

Figure 3.2: The ACT-R architecture: modules and assumed corresponding brain regions (ABB+04). The data buffers associated to the four modules are merged and activate so called productions in memory, aiming to solve the modules problem statement.

capabilities[6]. The main driver for this approach was Jerry Fodor, who argues that such modules must have specific properties (And07, p. 56ff.). Hence, different from other cognitive systems, where modules are chosen as an information-theoretic representation of the mind's function, i.e., aiming for functional equivalence, ACT-R has strong claims about the correspondence of its modules to brain modules. After years of developing the modular structure, ACT-R's modules are mapped to brain regions (rather than vice versa). Figure 3.2 shows the relation between the ACT-R modules and describes the mapping to assumed brain regions (in brackets).

The modules and capabilities in ACT-R focus on those cognitive capabilities that supposedly distinguish human from other animals (And07, p. 187ff.). Hence, ACT-R is mainly concerned with high-level cognition, and its prototypical application lies in algebraic problems. In particular ACT-R is used to replicate how children learn to solve equations. This should enable identifying learning problems and develop supportive strategies[7]. To compare ACT-R's performance with experimental data, the following measures are used: time to perform the task, accuracy in the

[6]Evolutionary psychology assumes that modules evolved because entities that were modular could respond to change more quickly, and therefore had an adaptive advantage.

[7]The insights of theses studies are embodied in a widely used cognitive tutoring systems

task, and (more recently) fMRI (functional magnetic resonance imaging) data. In the example of solving equations, ACT-R is provided with middle-school knowledge that serve as a pre-requisite for solving equations (And07, p. 21). After that ACT-R is run with the same algebraic problems as middle-school children (11-14 years), which are asked to learn to solve equations for six days. ACT-R was able to predict the childrens' required time to solve the equation for each of the six days (And07, p. 22). To be able to directly compare ACT-R's results to that of the children, the whole process - from perceiving an equation to entering the solution on a keyboard - is simulated. This is referred to as an end-to-end approach of validation (And07, p. 22).

Beside the main application to examine high-level cognition, other fields of applications are simulations of car-drivers (Sal06) and simulation of users in human-computer interaction (AFR05). Despite the distribution and popularity of ACT-R, it is almost not used for social simulations.

One of the rare cases is the usage in a game-theoretic approach to replicate how humans develop and maintain trust (KK13). Here a "voluntary trust game" is simulated, where the first player decides to take a sure (and equal) payoff (i.e., the other player does not have to decide) or let the other player decides how to split the payoffs. An iterated game is conducted with probands, and for different theories (explaining the participants behavior to trust or not) different sets of production rules are used. Typically a scale of 15 is used. The results show that a simple cognitive strategy is superior to a theory-of-mind strategy. However, besides the critic that the usage of game-theory to account for human behavior is only limited possible (see Chapter 3.1.1), as typical in ACT-R, the theory behind production rules is the sole explanatory concept. However, it can be argued that human behavior and decision making cannot be explained solely by reasoning capabilities. This is not so obvious when solving equations but more so when modeling trust. However, this example shows again the popularity and advantage of ACT-R for directly comparing the result and behavior with human data by replicating the behavioral process from perceiving the game on a screen until entering the decision with a keyboard (cf. end-to-end analysis, see above).

Another recent example of using ACT-R in social simulations is the application for simulating collective sense-making (SS14). The ELICIT (Experimental Laboratory for Investigating Collaboration, Information Sharing and Trust) sense making task is used, where participants are shown a subset of information that is required to decide a situation (e.g., if a terrorist attack will occur). The researchers define two kinds of chunks: a statement (consisting of object, attribute, and value, similar to a triple in RDF (Resource Description Framework)) and a message (consisting of the source, the target, and the content). To be able to decide the task, 124 production rules are formulated based on these chunk-types. This again shows that every decision factor must be represented in production rules to be considered, i.e., it is the central bottle-neck and the sole mechanism for decision making.

The ACT-R architecture is extended with a message module (with one buffer for sending and one for receiving messages) and various lisp-databases to provide communication functionality between the agents. The databases are used for centralized communication between the agents (hence, it is not a typical multi-agent system); they specify the communication network of the agents and mediate the messages between them.

Obviously, the usage of ACT-R in this model is not meant to represent human decision making realistically, since "... the particular strategy adopted by agents to solve the ELICIT sensemaking task is unlikely to be the same as that used as human subjects." (SS14, p. 203).

3.2.3 CLARION: A Dual Processing Cognitive Architecture

CLARION (Connectionist Learning with Adaptive Rule Induction Online) is described as an integrative modular architecture of dual processing (Sun09), aiming to tackle the interaction between implicit-explicit knowledge, between cognition and metacognition, and between cognition and motivation, as aspects that are neglected in conventional cognitive architectures such as SOAR and ACT-R (see above) (Sun06a). CLARION consists of the following interacting components (see Fig. 3.3): the action-centered subsystem (ACS), the non-action-centered system (NACS), the motivational system (MS), and the meta-cognitive system (MCS). The ACS controls external actions and internal (mental) operations, the NACS maintains general implicit and explicit knowledge, the MS is concerned with providing feedback about how satisfactory outcomes were, the MCS monitors and possibly modifies the operations of the other subsystems.

Following a dual-process theory of mind, every subsystem consists of a bottom-level that processes implicit sub-symbolic knowledge, and a top-level that processes explicit symbolic knowledge. On the one hand explicit knowledge as innate knowledge guides the formation of implicit knowledge, considered as top-down learning. On the other hand explicit knowledge is extracted in bottom-up learning. The formation of explicit knowledge serves the availability of better (since directly) accessible and manipulable knowledge. However, the dominant way of learning is done bottom-up, with top-down learning being primarily relevant only at the beginning of learning.

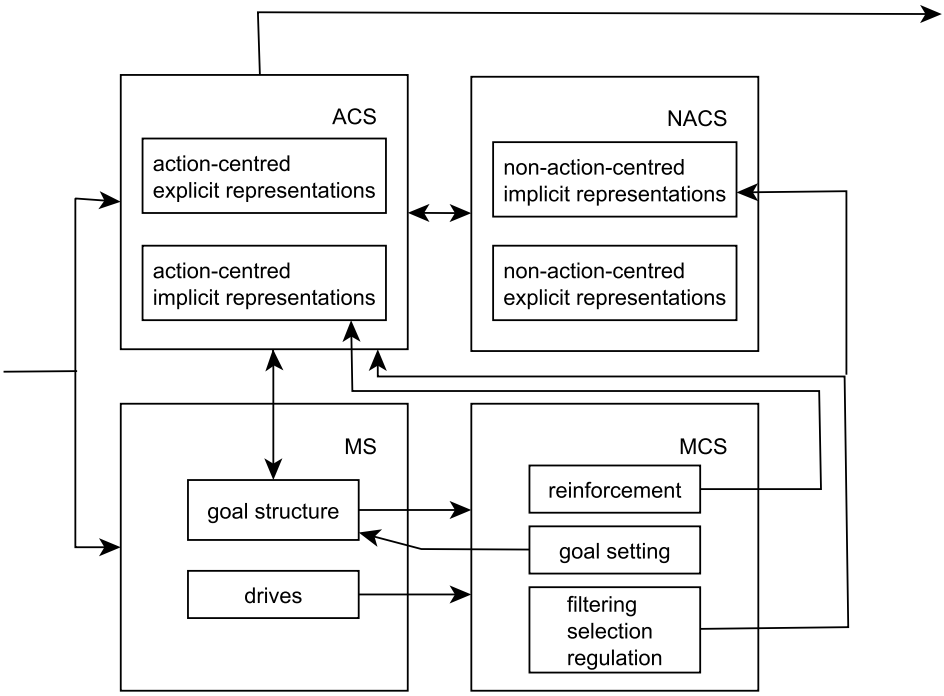

Figure 3.3: The CLARION architecture (Sun09). Modules are highly interconnected, with the ACS getting input from and generating output to the environment

The action-centered subsystem is the central part of CLARION (Sun06a, p. 84), concerned with action decision making. As all subsystems, the ACS is distinguished in a bottom-level and a top-level. The first consists of neural networks that evaluate all possible actions regarding how desirable it is in the given situation (given by sensory input and the current drives in the MS),

using the reinforcement-learning algorithm 'Q-learning' (Sun06a). The top-level considers explicit knowledge in a rule-based system, where generalized rules about actions are created based on the learning success at the bottom-level, using a Rule-Extraction-Refinement algorithm (Sun06a). In decision making, if appropriate, the rules of the generalized top-level are used, otherwise the bottom-up level goes for trial and error.

The NACS is controlled by the ACS. It is mainly concerned with forming and fetching knowledge about the world. At the bottom-level it creates associative memories based on learned associations between input and output data. Based on them, explicit associative rules are formed in the top-level, among others by similarity-based reasoning and making inferences about the world.

Opposed to many other cognitive architectures CLARION considers a motivational system. It uses Hull's concept of drives (Hul51) to represent motivations. A generalized notion of drive is used, were drives represent "... internally felt needs of all kinds that likely may lead to corresponding behaviors ..." (Hul51). Primary and secondary drives are distinguished. The former are separated in low-level primary drives, mostly concerned with innate physiological needs (such as hunger), and high-level primary drives (such as affiliation and belongingness), mostly concerned with social needs. Secondary drives are derived from primary drives and are learned by conditioning (e.g., the pursuit of money). A clear distinction to other concepts of drives is the determination of the drives' strength, which in CLARION is determined by an internal deficit *and* an external stimulus (e.g., food in case of hunger). The latter determinant emphasizes Hull's behaviorist approach and shows a central difference to a psychoanalytical approach to drives, where the inner world (e.g., the bodily deficit) is the sole basis for determining a drive's strength. Overall, the separation of different criteria in processing drives, e.g., reflected in bottom-level and top-level, is not clear, in particular how such aspects as contiguity of actions and persistence (Sun06a, p. 91) are necessary in the MS.

Amongst others, CLARION is used to explore organizational decision making (NS06b). Replicating simulations, where the impact of organizational structure and information access is explored (CPL98), CLARION is used to draw further conclusions by experimenting with learning behavior and varying parameters regulating cognitive capacities. The organizational decision making task is to cooperate on deciding if a radar signal is a hostile aircraft or not, given nine attributes of the aircraft. This task is originally simulated with two organizational structures: (1) teams with autonomous agents (where the organizational decision is determined by the majority), or (2) hierarchies with a chain of command; and two types information access: (1) distributed access (where agents only see a subset of information) , or (2) blocked access (where all agents have access to the information). Simulations with CLARION are able to replicate the findings that team-work outperforms hierarchical structures and distributed information access proved superior to blocked information process. Using CLARION instead of simplified agents in (CPL98) resulted in a better coverage of the human data from experiments. Additionally, simulations with CLARION demonstrate the impact of the learning duration on the performance of different organizational structures and information access types, and the impact of different kinds of processing and learning in CLARION (see above) (NS06b). In particular, it is shown that for different possible configurations of CLARION agents, different organizational structures and information access are superior. For instance, considering the learning duration, the superiority of team organizations and distributed information access is only valid at the beginning of learning. Also, experiments show the impact of the learning rate, parameters that regulate the generalization and application of rules, and parameters that determine which level (bottom or top) preferably to choose. Additionally, some experiments show that varying cognitive parameters do not have an impact on the original outcome. These conclusions enable fine-grained recommendations for

organizations. For instance, the findings in (NS06b) could be interpreted as recommending the assignment of organizational roles depending on the individual cognitive capabilities, e.g. learning capabilities. Additionally, the findings could advise organizational policies, e.g., that rule-based learning should be used at the beginning of new projects.

Such simulations show that using generic, instead of task-specific, assumptions and parameters, i.e., to widen the distance between assumptions and outcome (NS06b), enable deeper explanations and more flexible experiments. However, this is not the case, if cognitive parameters do not have an impact on the outcome, or if their impact is representable abstractly. This may be the case in another simulation application of CLARION, where the number of authors that submit a specific number of papers to a scientific journal should be predicted (NS06a). In this case the CLARION-simulation was not able to outperform other simulations with simplified agents. However, as the authors argue, these simplified simulations used assumptions, which were developed to match the human data, which is not the case in CLARION's the generic assumptions.

3.2.4 PSI: A Motivated Cognitive System

The PSI theory (referring to the greek letter, which is used as an abbreviation for Psychology (DG13)) claims to be a blueprint for *a* mind (Doe99) (Bac09, p. 54), and sets the focus as an architecture of motivated cognition. However, the PSI theory does not claim to represent mechanisms of human cognition (Bac09, p. 311), but rather it aims to generate behavior "... of those classes that we would call cognitive" (Bac09, p. 311). Even if inspired by human cognition "... PSI agents are animals of a kind different from humans" (Bac09, p. 155). This approach is also termed as cognitive AI (Bac11). However, Doerner's original objective is to use simulation as a psychological methodology. That is, to follow an engineer-based approach and build a machine that shows the capabilites of the human mind and thus develop a candidate theory of the human mind [8]. Hence, PSI is described originally as "a formalized computational architecture of human psychological processes" (DG13, p. 297), recently relating the functionalities of the PSI theory to brain regions (DG13, p. 299).

The main theme of the PSI theory is how motivations are used for action regulation. "... the PSI theory's unique contribution to cognitive science is the way it combines grounded neuro-symbolic representations with a poly-thematic motivational system. It is offering a conceptual explanation for both the multitude of goals a mind sets, pursues, and abandons during its cogitation and action, and the perhaps equally important serendipity of wandering thoughts and associations. By including an understanding of *modulated* cognition that treats affective as particular configurations of perceptual processes, action regulation, planning, and memory access, the PSI theory offers a non-trivial integration of emotion into its architecture, consistent not only with external observables but also with the phenomenology of feeling and emotion"[9] (Bac09, p.303). This is done by a unified parsimonious model that tackles these aspects (only) implicitly. Hence, PSI follows a holistic approach in developing a model that is able to explain the human mind. This could only be done following a mechanistic, rule-based approach, which is seen as a criteria to regard psychology as a natural-scientific discipline (Bac09, p.53) [10]. Inspired by cybernetics,

[8] "Das Ziel unserer Arbeit ist es, das Gehirn im Computer nachzubauen, aber im Mittelpunkt steht dabei nicht so sehr die neuroanatomische Struktur, sondern die Funktion." (https://www.researchgate.net/publication/220633938$_{I}nterview_{m}it_{D}ietrich_{D}orner$)

[9] Again, this contradicts the description as cognitive AI.

[10] Here Dörner is inspired by Aristotle, who declared "The soul is the principle of the living", which is interpreted as "... the soul is the set of rules that determine the functioning of an organism, if it is alive" (Bac09, p.53).

humans are regarded as self-regulated mechanistic systems, with the homeostatic principle valid on the physiological and psychological level.

Sense-think-act cycle and concepts

The focus and the trigger of cognitive processing in PSI agents are pre-defined demands that may give rise to motivations via urge-signals (Bac09, p.69). Demands are physiological (energy, water, integrity), cognitive (certainty, competence), or social (affiliation) see Figure 3.4. Actions that fulfill demands are indicated by a pleasure signal (decreasing the demand), aversive actions with an displeasure signal (increasing the demand), which are subsequently used as reinforcement signals. An unsatisfied demand is indicated by an urge (i.e., the difference between a setpoint and the current value) and selected as the currently active motive[11], which is represented as a demand and satisfying goals, dependent on its chance of satisfaction. The selected motive, then, pre-activates experienced situations in associative long-term memory that have satisfied the demand, which are processed as goals, with activating action sequences to get to the goal state (i.e., stored plans) if possible, routine behavioral patterns (schemas (DG13)) if available, constructing new plans otherwise - using hill-climbing and an activation-based search to reduce the distance between the current situation and the goal situation, or as a last possibility follow a trial-and-error approach [12] (DG13, p. 299) (Bac09, p. 70). This process implicitly activates object and episodic schemas that influence the perception (and expectation[13]) of external stimuli, forming a situation image. The current image is associated to the previous and subsequent situation image in a protocol chain, with the association strength depending the impact (pleasure, displeasure) the situation had on an agent's demands (Bac09, p. 70). The protocol, then, is stored abstractly (or reinforced) and can be used as a future plan.

Modulators (see Figure 3.4) enable the adaption of the cognitive processes to the current needs (Bac09, p. 72). Strong urges decrease deliberation and increase action readiness through an activation modulator of corresponding processes, which in turn influences a resolution modulator, impacting the resolution of the schemas used in perception, memory organization and retrieval, and deliberation (Bac09, p. 73). The third modulator - selective threshold - is added to the currently active motive's relative strength and thus modulates the readiness to change motives. Overall, modulators control behavioral tendencies and represent an implicit and emergent concept of emotion.

Hence, different configurations of these modulators can be described as emotions. For instance, anger corresponds to high activation modulation, high selection threshold, and low resolution level (DLD$^+$01). This constellation could occur, if an agent cannot reach its goal. In such situation an agent would have a high action tendency and risky behavior. Another example can be given by anxiety, which is represented by a high level of activation, low level of resolution, low level of certainty and competence (DG13, p. 311). In such configuration, the PSI agent has low expectations, jumps between motives, and has a low level of deliberation.

In this sense emotions cannot be described as system states, but as flows of motivational and cognitive processing (DG13, p. 309). Nevertheless, this approach is mapped to a dimensional representation of emotion. For instance, Wundt's pleasure-displeasure dimension (see Chapter 3.3) is represented by reinforcement signals of PSI's competence demand, the excitement-inhibition dimension corresponds to the competence demand's level, and the excitement–inhibition dimension is represented by the activation modulator (DG13, p. 308). Overall, the function of emotion

[11] Only one motive is followed at a single time.
[12] This process follows the Rasmussen ladder (Ras83) (Bac09, p. 310).
[13] Unexpected events increase uncertainty, see Figure 3.4.

is regarded as the adjustment of an agent's behavior to the demands of the current situation, where "... motivations determine what must be done, emotions determine how it is to be done ..." (DG13, p. 307).

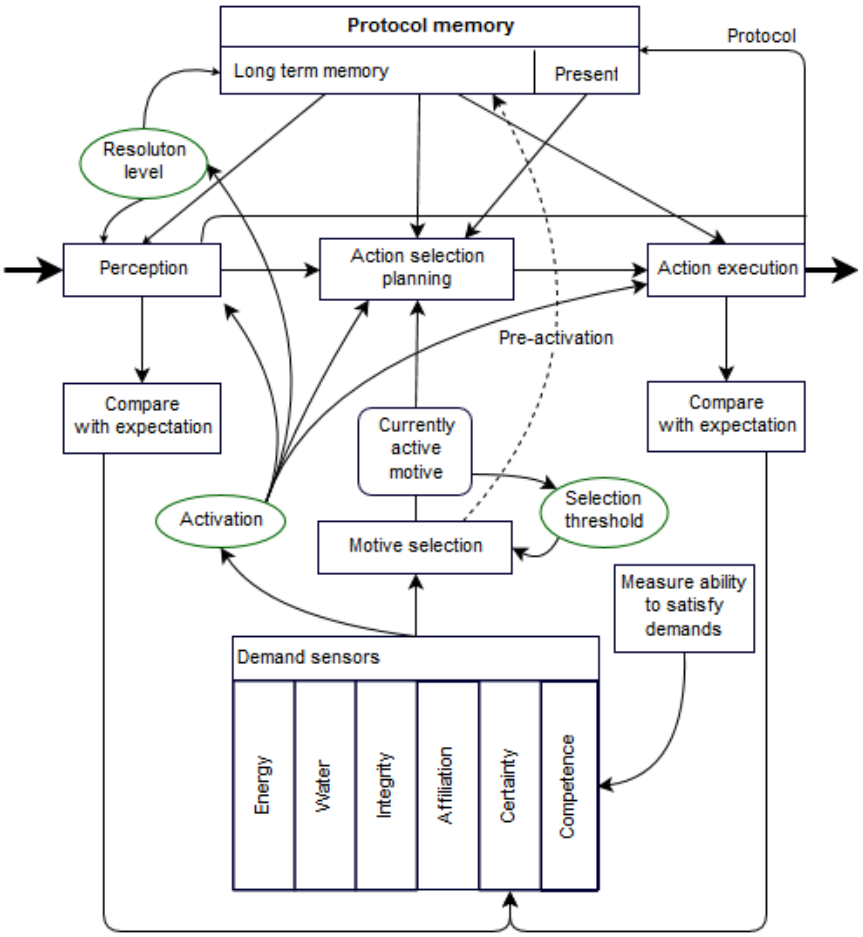

Figure 3.4: PSI architecture (modulators in green) (Bac11, p. 73). Pre-defined demands generate motivations that regulate behavior directly through associations to action plans (together with perceptions) and indirectly through modulation of this process.

Building blocks

The PSI theory uses neuro-symbols as basic building blocks and form hierarchical node networks for declarative, procedural, and tacit knowledge, with special circuit nodes that regulate the formation and solving of associations and spreading of activation (Bac11, p. 304). However, they are not modeled after actual brain structures, since these are not considered as crucially for understanding cognition[14] (Bac11, p. 62).

Evaluation

To test the PSI theory of the mind mainly Artificial-Life-Simulations are used to demonstrate that PSI agents are able to adapt on different environments and solve environment-specific problems. The prototypical example was to populate a virtual island with the PSI-steam-engines 'James'

[14]To wait for neuro-science to have all necessary insights is even seen as hindering researching the function of the mind (Bac11, p. 73).

(DLD+01), whose job is to gather 'nucleotides' for power plants. To survive the agent needs water (from lakes) and hazelnuts or sunflowers. In foraging for food and nucleotides James has to avoid dangers to avoid pain. The agent learns concrete goals (where and how to get food) in exploring the island. Besides the physiological demands, cognitive demands are pre-defined: the demand for certainty about the environment and the demand for competence in acting upon the environment. In particular, certainty is obtained if James is able to anticipate the environment, the effects of his actions upon it, and the applicability of his plans; competence is obtained when actions lead to the (expected) satisfaction of his demands.

The next step of testing the PSI theory was to use such Artificial-Life-Simulations as a computer game in experiments. Human probands have to play e.g., James and take care of his survival (BD98). This behavior is then compared with the simulation of autonomous PSI-steam-engines, which led to insights how to adapt the PSI-theory to account for better human-like behavior (DG13).

Another way of validating the PSI-theory is to test whether it is able to predict known reactions of humans in specific conditions. Examples therefore are the reaction to the availability of alcohol (HD09)(DG13). Another example is to compare the crowding behavior of virtual mice (without explicitly accounting for crowding), which live under similar conditions as James (see above), with known human crowding behavior. (DG13) emphasizes particularly on the similarity between emotional behavior and experience in humans and simulated PSI-mice in crowd conditions.

3.2.5 Consumat: a Socio-Cognitive Agent Model

The development of cognitive architectures and agent-based social simulations evolved independent from each other. The former lack to consider the social dependencies of cognitive functionalities. The latter lack to consider psychologically plausible assumptions about their abstract models of decision making. Recently different approaches in ABSS and cognitive science (e.g., Jager 2003; Sun 2006) argue against the usage of simplified, psychologically implausible decision models in ABSS. Typical arguments are that many macro-level phenomena are caused by the dynamics between the micro and macro level, which is not captured by (too) simplified models (Jager 2003).

One of the evolved agent models in ABSS that claims to account for psychologically realistic behavior is the Consumat model, which aims to go beyond the typical models of homo economicus towards a model of homo psychologicus (JJV+00). The approach provides a conceptual framework to include different human needs and decision strategies. Existence, social and personality needs are modeled as the main drivers of the agent. Existential needs are regarded as the need for food, income, and housing. Social needs consist of the urge to interact with others, to belong to a group, and to have a social status. Personality needs are satisfied by following norms and values (JJV+00). Based on the (current) relevance of these needs, an agent chooses appropriate behavior to satisfy them. The execution of the available behavior options also depend on an agent's abilities (e.g., income, land possession, availability of tools, cognitive capacity). Besides the state of satisfaction of an agent's needs, the agent's uncertainty determine which type of decision strategies are chosen: repetition (in case of low uncertainty and high satisfaction), imitation (high uncertainty and high satisfaction), deliberation (certain but dissatisfied) or inquiring with others (in case of uncertainty and dissatisfaction). The selection of decision strategies is also reflected in the interactions between the agents: social decision-making may focus on a rather normative influence (imitation) or informative (inquiring). Memories about behavioral opportunities and

other agents' behavior and abilities support to decide how to apply the decision strategies which is only updated if cognitively demanding strategies are being used.

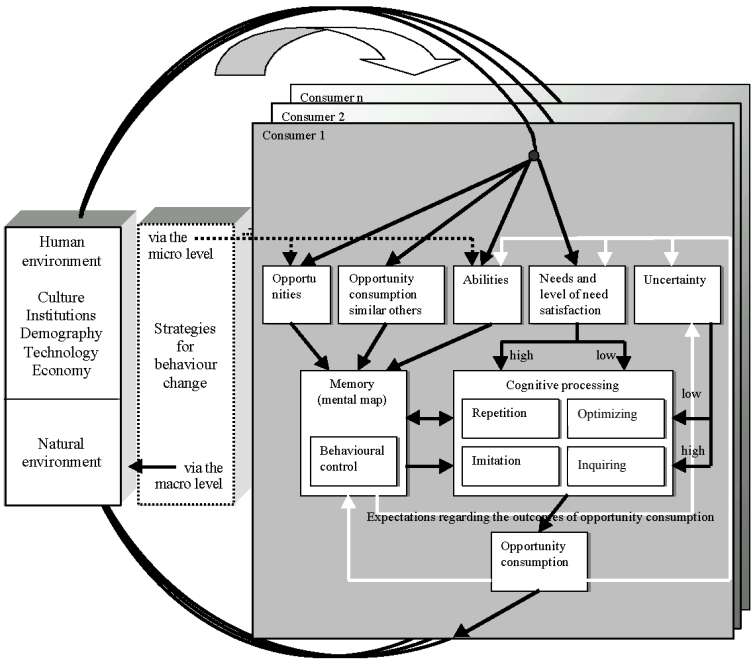

Figure 3.5: Overview of the Consumat framework (JJV+00). Agents select decision strategies (repetition, imitation, optimizing, inquiring) based on their uncertainty and needs. Opportunities and the agent's abilities determine which memories to use in this process.

The Consumat approach is used in various domains of consumer decisions of sustainable products. Examples include the diffusion of environmentally friendly products (Jl02), simulation of sustainable life styles (BVCP13) and sustainable transport choices (NB14). Here the Consumat approach shows how the consideration of different decision strategies is able to explain dynamics in sustainable behavior.

Comparing and Connecting Consumat and SiMA-C[15]

The commonalities between Consumat and SiMA-C (see Chapter 5) show the importance of considering a generative approach to motivations and different modes of decision-making. This is especially important when the motives, and the external (environment) and internal (personality, agent's state) conditions of sustainable behavior are at stake. Hence, in explaining sustainable behavior, both, psychological aspects (such as different demands from needs and norms) and social aspects (such as social pressure) and their integration have to be considered. The Consumat and SiMA-C approach are both able to provide a framework to integrate psychological and sociological perspectives and enable to explain behavioral dynamics by the interplay of multiple model variables. Consumat and SiMA-C are similar in basing behavior on motivations and in considering heuristics. The Consumat integrates these aspects more explicitly. For instance, a core process addressed in the Consumat framework – the selection of a suitable decision strategy

[15]This section is taken from a part of (SJD16) that is written by the author of this thesis

given the situation – is not explicitly considered in SiMA-C. However, as general in the generative-functional SiMA-C model, the resulting interaction between its components is able to generate similar patterns of decision making as in Consumat. SiMA-C focusses on how these different strategies are generated by underlying mechanisms, most importantly emotion, a driver not explicitly addressed in the Consumat. Also within SiMA-C a more explicit modeling of the memory is formalized, allowing for the parameterizing of storage and retrieval of concrete memories. The higher degree of parametrization and memory-processing (e.g. spreading of activation) enables more detailed model exploration and experimentation, while it makes simulations (unnecessarily) more complicated. Therefore, we expect that the Consumat model is more efficient in modeling the basic behavioral dynamics of a population that is interacting within an environmental relevant setting (e.g. resource use, consumption), whereas SiMA-C offers a possibility to experiment with strategies addressing the emotions and experiences of this population.

The compliance between the Consumat and the SiMA-C approach, and the difference in focus and level of details, provide possibilities to join the two approaches for deeper explorations of social phenomena. As Consumat provides a lightweight model, practical value in social simulations, and considers key factors of a psychological model, it can be used as a starting point to identify cases in simulations that require further exploration (e.g., because it identifies a wide variety of possible outcomes, or unexpected social barriers for change). In such cases SiMA-C may be able to provide more insights of agents' decision-making processes in different personalities and further specify Consumat's variables where (assumedly) necessary. For instance, using a model of emotion in SiMA-C (and the dynamics of memory activation and valuation), detailed reasons for why agents choose goals can be provided. This also enables fine-grained parametrization with empirical data. Overall, Consumat enables a platform for broad policy testing, and may direct further explorations with SiMA-C, when it is assumed that the focus of SiMA-C provides detailed variables (for policies) to dock on. This would also support individually tailored policies and simulation experiments that address particular key-agents (e.g. opinion leaders) in a more precise way.

3.2.6 Simulation of the Mental Apparatus and Applications (SiMA)

The SiMA approach[16] was initiated by Dietmar Dietrich (DS00) (DFZB09) due to recognizing the lack in artificial intelligence research to solve everyday problems. Projects concerned with recognizing dangerous situations in kitchens (SRF00) and airports (BKVH08) have demonstrated that state-of-the-art technologies are incapable of solving these problems sufficiently. Hence, conventional approaches in AI have not led to artificial systems that are able to cope with complex tasks in daily human life that humans are able to handle in an automatic, i.e., unconscious, fashion. In the SiMA project complex systems are regarded as entities that cannot be described completely due to missing information (DSS+16). The SiMA approach, then, aims to model the human mind and explore how it deals with complex tasks (DSS+16). Instead of focusing on reasoning capabilities, artificial intelligence system that aim to solve every-day problems have to consider the unconscious capabilities of humans. That is, 'intelligence' does not solely lie in reasoning capabilities, but is based on human capabilities that can be regarded as intuitive. This insight led Dietrich and his research group to go back to basic research in artificial intelligence and ask an instance of the original question in AI: Which capabilities allow humans to solve everyday problems and how can we harness them in artificial systems? In doing so, intensive collaboration were

[16]The SiMA project was called ARS (Artificial Recognition System) project until 11.02.2015.

done with researchers from various disciplines, e.g., neuro-scientists, psychoanalysts and other psychologists. This was regarded to be necessary instead of finding computational solutions that solely follow criteria of computer science. The role of computer technology in this endeavor is to translate theories about human information processing into a computational model. Following the experience in dealing with information-processing systems some principles have to be considered in such knowledge translation and the models from other disciplines have to be compatible with these principles for the knowledge translation to be successful.

3.2.6.1 SiMA Approach

The premise of developing systems with human-like capabilities is to find theories and translate them to a computational model of human information processing. Such a model would provide insights about the required capabilities to solve problems in a human-like way. In turn, this provides the required information to build artificial systems with similar capabilities. Hence, the first step was to evaluate models of human thinking regarding their applicability for the SiMA approach. This analysis resulted in the selection of psychoanalytical metapsychology[17] as the sole theory that fulfills the principles of the SiMA approach.

Some specific key features define the SiMA approach. First and foremost, it aims for a functional model, i.e., it follows a generative approach with the focus on describing functions that generate behavior instead of building a behavior model. This enables a generic and flexible model. Second, a key feature is the layered description of human information processing. The principle here is to use appropriate means of description for different aspects of a systems, e.g. the neuronal layer should be described with other means than the psychic layer. Third, in developing such a model a holistic and unified approach is used, which considers a consistent and coherent description of all key aspects of human information processing. However, 'holistic' does not imply to include all mental functions in detail, but to model the flow of information holistically, i.e., considering all steps from sensors to actuators (DSS+16). The consideration of these key features is only possible by following a bionic and hence interdisciplinary approach.

A Layered Approach to the Research Object

The usage of a layered approach is the key idea from computer technology that enables a holistic description of the human brain's functioning. This separation of its description in layers follows a common methodology in computer science (cf. the ISO/OSI model) and supports comprehensive modeling. Naturally, each perspective has its own form of description. One would not, for instance, describe a text-processing system by describing the transistors of the underlying computer machine (DSB+13b). Similarly, it is not feasible to use a neuronal model to describe psychic capabilities directly. The SiMA approach uses three layers to describe human information processing (see Figure 3.6).

The first layer, representing neurons, is described as hardware under consideration of the laws of physics. The next layer, the neurosymbolic layer, is concerend with the symbolization of the

[17] Many consider the scientific (opposed to the clinical) branch of psychoanalysis as the most appropriate functional and holistic theory of the human mind (see e.g. (Kan99, p. 505)) and call for more integration of psychoanalytic insights into neuroscience (e.g. (Dam12)). Although the perspective of psychoanalysis as a scientific discipline is of debate (see e.g., (War97), (Gar86)), most of its basic theories (e.g., the principles of homeostatic systems and drives, the relevance of memories, the role of the unconscious etc.) is widely accepted. Also, in formulating testable assumptions based on psychoanalytic theories by the SiMA approach, scientific evaluation is supported by a computational approach.

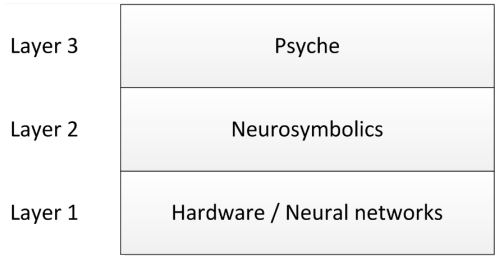

Figure 3.6: Layers in the SiMA model (DSB+13b). Different perspectives on the human information-processing system has to be described by different means.

neural layer, i.e., it bridges the physical and information domain. The third layer represents the mental layer, which is described in functional terms in an algorithmic form. It is important to emphasize that these layers in the end describe the same entity. Hence, psyche and brain are (of course) considered to be the same entity; they only represent different perspectives on the object of research (Forschungsgegenstand).

The object of research is the information-processing system of humans, called information organism (see Figure 3.7). There are various approaches to gain knowledge about this object of research, with different premises - reflected in different research material. In the SiMA project two classical perspectives on the information organism are considered and integrated. The neurological approach aims an objective view on the information organ using physical data (e.g., EEG (electroencephalogram), fMRI (functional magnetic resonance imaging)) as research material to represent the information organ as a biological model (i.e., the brain). The psychological approach aims to develop subjective models using data about human behavior as research material. Hence, this approach uses the output of the information organ as research material to represent it in psychological models. Different approaches, such as neuropsychoanalysis, aim to map these two perspectives (the neurological and psychoanalytical perspective). As usual in neuroscience, such correlations are usually done on a statistical basis. Instead, the SiMA approach provides a method for a causal relation between the two perspectives by introducing an additional, neuro-symbolic, layer and specifying the interfaces between the three layers. Since only a functional description is relevant for artificial systems (but not how the functions are implemented), the SiMA project focuses on the third layer, i.e., the description of the mental apparatus.

Metapsychology as a Framework Theory for Layer 3

Which model of the mental apparatus is appropriate for the functional and holistic approach sketched above? The key criteria for the usage of such a psychological model in an interdisciplinary methodology to develop a computational model of the mind are as follows. A top-down approach requires a holistic framework theory, considering both unconscious and conscious aspects, that guides the top-down process in its specification of mental functions. Such a framework solves a common problem in cognitive architectures by enabling a holistic approach to develop an integrated model. In particular, it serves as a glue that guarantees a coherent and consistent model. Hence, the psychological model should be appropriate for the use as a framework and guideline for developing an integrated functional model of the human mind. Another requirement stems from the functional approach in the SiMA project, which prevents mixing correlation and causation. Hence, the psychological theory has to describe how behavior is generated and caused.

As the most suitable model that fulfills these criteria, and is thus appropriate as a framework, the second topographical model of Psychoanalysis is chosen (DSB+13b) (see above). It uses

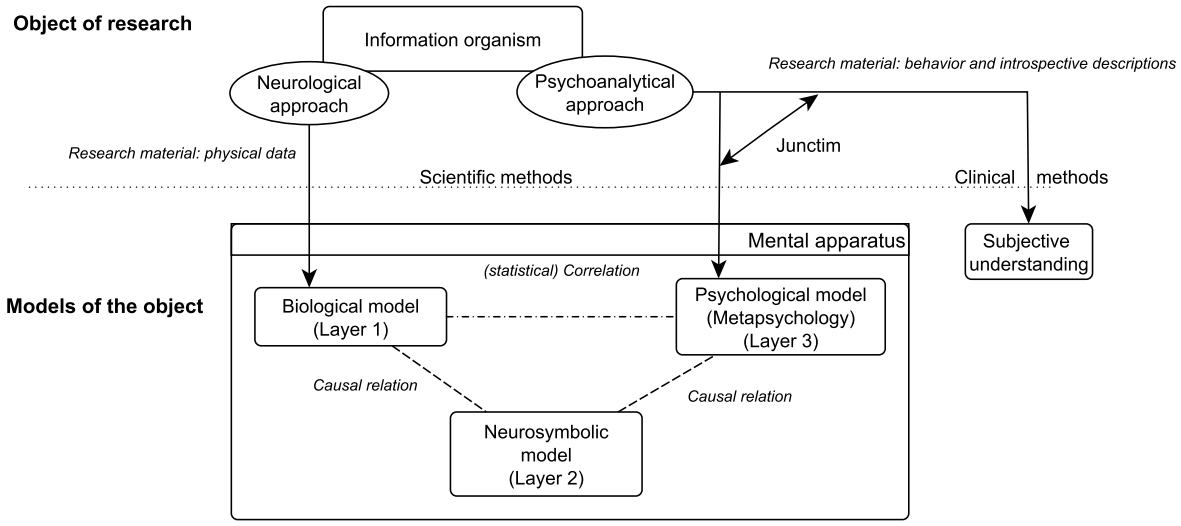

Figure 3.7: SiMA research object and its models. Disciplines use different accesses to the research object (the information organism) that use different gained material to build models, which represent different perspectives on the information organism. Riding these different perspectives by using a layered approach (cf. Fig. 3.6) is a premise for a holistic understanding of the research object.

the abstract functions Id, Ego and Super-Ego to describe the human mental apparatus (for an overview see Chapter 4.2.1). The Id represents drives, which are in effect bodily demands coming from internal sensors, the Super-ego represents internalized moral demands and the Ego mediates between the Id and the Super-Ego under consideration of the external environment (BDZ+13).

The framework characteristic of the second topographical model is also reflected by terming the underlying theory meta-psychology. However, such a framework can only provide an abstract model and a point of departure for model specification. Using the second topographical model as a framework, it must be specified and complemented with established findings from neuroscience, psychology and similar sciences, as long as there are no contradictions. For instance, Damasio's (Dam03) concept of emotions and feelings supports the specification of how motivations and emotions are related (SHDD14). In interdisciplinary collaboration with researcher from different disciplines, who often do not follow a strict deterministic approach, some difficulties arise which impede the direct usage of psychoanalytic knowledge for a technical model. This reflects often stated criticism of the use of metapsychology, in particular its abstract character. However, such difficulties are tackled by specifying the framework by established findings using an axiomatic approach. Such an approach forces psychoanalysts and neuroscientists to sharpen their concepts and models until they are appropriate for translation and specification into a causal and deterministic functional model.

3.2.6.2 SiMA Model Overview

The three abstract functions of the second topographical model are specified incrementally until their description can be used for an implementable functional model. The most concrete level of the SiMA model is shown in Figure 3.8.

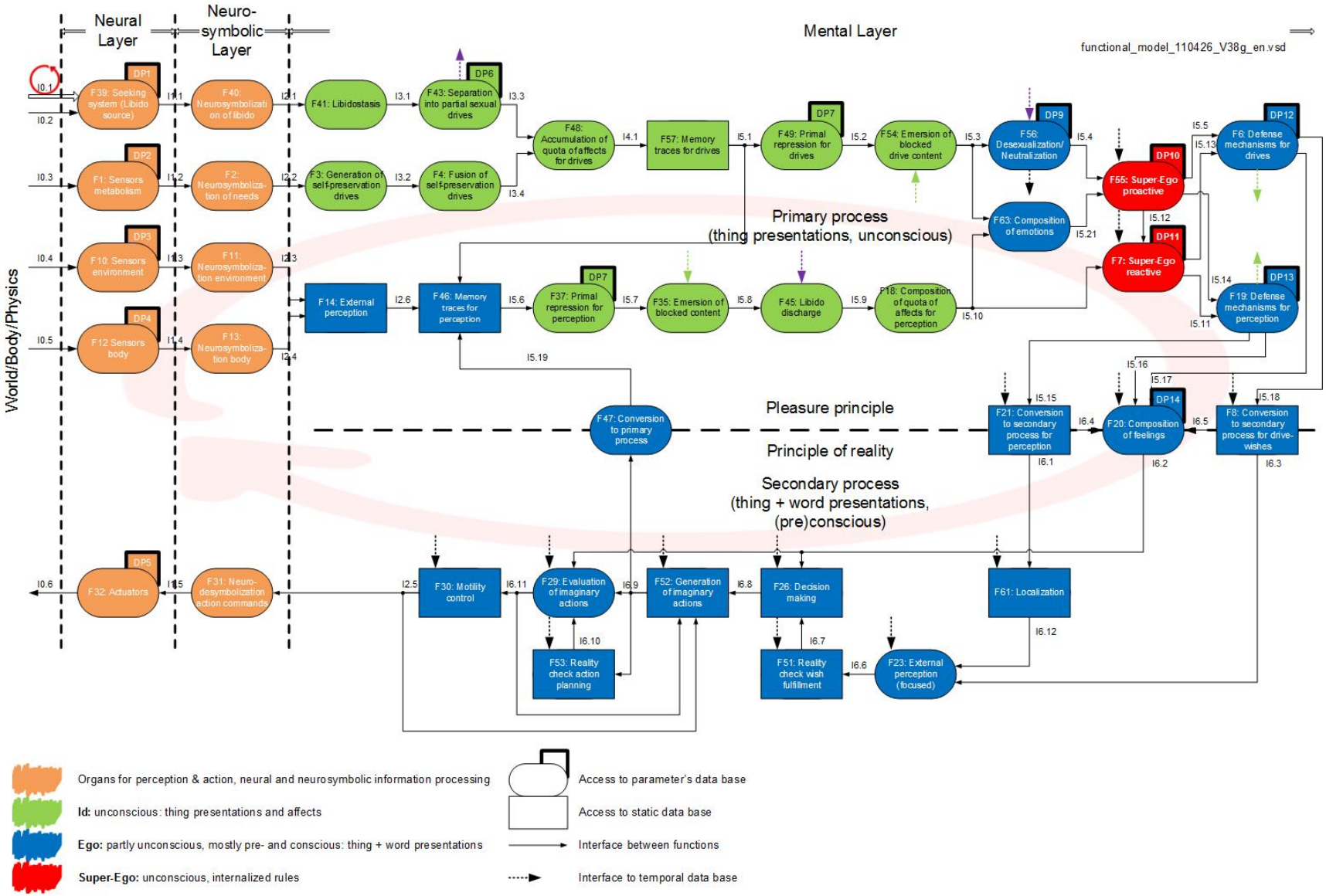

Figure 3.8: SiMA functional model Version 38g.

However, for a model overview a track view, where coherent function modules are summarized into tracks, is used (see Fig. 3.9). The functional model has four inputs, two for drives and two for perception, and an output for the decided actions. Drives represent the agent's desires stemming from bodily needs. Perception is distinguished in self-perception and environment perception. Drives and perception are compared and associated with the agent's experience to give a basis how the system may satisfy its desires in the perceived external world. In the defence track, desires may lead to conflicts with internalized rules. The conflicts are handled by so called defence mechanisms, which may transform desires to social accepted ones. Next, goals are generated from desires and the affordances of the external environment. Finally, the best evaluated goal is chosen, a plan and action is selected that satisfies the different demands in the current situation best. In the action track this action is finally executed.

As indicated by the horizontal dashed line in Fig. 3.8, the SiMA model distinguishes two modes of processing: the primary and secondary process. The primary process follows the pleasure principle, aiming for immediate and maximal satisfaction of the agent's needs. The secondary process extends this principle by considering the affordances and limitations of the external reality. The two processing modes also differ in the processed data structures. Opposed to the secondary process, the primary process does not consider hierarchical (including temporal) relations between data.

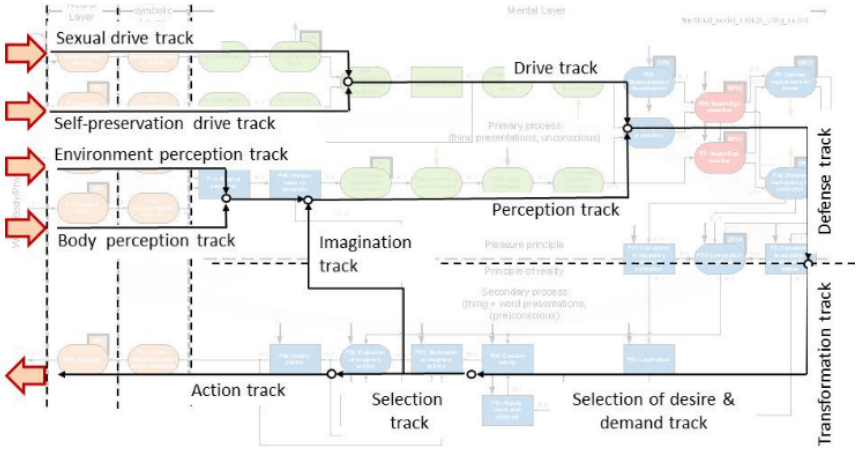

Figure 3.9: SiMA functional model Version 38g - track view.

A central part of the SiMA model is considered with the generation of the agent's agenda and how the agent may cope with the external world in pursuing this agenda. Solving this question is a premise for any autonomous system and the basis of decision making. In SiMA human-inspired valuation models are harnessed to generate and prioritize goals as a basis for decision making. These evaluations occur incrementally under different principles and influences. Such a multi-level valuation model (SDD14), in which different valuation criteria are considered, shows how human-inspired valuation supports decision making by enabling an incremental determination of the relevance of goals and plans to reach them.

Next, those tracks that are relevant for the thesis at hand are described.

Drive Track

The drive track processes homeostatic data from the organs. The concept of a drive is a theoretical construct and is the psychic representation of a bodily need. The drive comprises a drive source, a drive object and a drive aim (DBD+15, pp. 85). The drive source is the organ, which

signals the bodily need; the drive aim is the satisfaction of the drive by an act; the drive object is the object used in this act. Two kinds of drives are distinguished: aggressive and libidinous ones. Every drive source triggers one of each kind. An aggressive drive is satisfied by an aggressive drive aim, a libidinous drive by a libidinous drive aim. After the generation of a drive by a bodily need, the drive is rated by a quota of affect, which reflects the drive tension, i.e. the tension of the drive source. The higher the bodily need, the higher the quota of affect. The next step in the drive track, called hallucinatory wish fulfillment, is the search for possibilities to satisfy the drive, i.e. appropriate drive objects and drive aims. The agent uses its memory to determine drive objects and drive aims that satisfied a similar drive according to the agent's experience. For further details see Chapter 4.3.2.

In this regard, the drive track represents the central part of the agent's motivational system, which in further parts of the SiMA model leads to desires and actions. The concept of the drive system represents the influence of the body on the psychic apparatus. In this regard, embodiment is considered as a central factor of the SiMA model.

Perception Track
The initial task of the perception track is the conversion of sensor data to symbolic data that can be processed in the SiMA mental layer. Therefore it merges sensor signals to semantic symbols in multiple steps. First, the raw sensor data have to be converted into symbols in the sensor interface of the SiMA agent. This is done by using the concept of neurosymbolization (Vel08). This process is inspired by neuroscientific principles. It uses multiple layers of so-called neurosymbols, which are the basic information processing units in neurosymbolization, and reflect features of neurons and symbols (Vel08). That is, every node in a neurosymbolic net has a symbolic meaning. In this regard neurosymbols represent a layer between neural networks - the hardware layer - and symbolic representation. The neurosymbolization layer considers multi-modal perception and merges sensor information from different modalities (related to the human five senses) to a multi-modal symbol by using a hierarchical concept of sensor fusion. As part of this process the binding problem is considered, i.e. all perceived information are assigned to a specific item. The result of the overall process is a symbolic item that is associated with its symbolic features, e.g., 'red', 'round', 'shiny').

The next step in the perception track is to categorize and recognize the result of neurosymbolization, i.e. a symbol. In the scope of the SiMA project this is done by a memory-based approach. Hence, the unknown symbol is compared to the agent's memory to identify and categorize it. As the SiMA agent represents a thirty year old person, it is assumed that the agent recognize all perceived objects based on its memory and does not perceive completely unknown objects. After identifying all perceived objects, they are used to recognize the external situation and activate similar situations (represented as memorized images). After that they are forwarded to the subsequent functional modules of the SiMA model.

Defense Track
A central aspect of the SiMA models are conflicts and their mediation by defense mechanims. Conflicts between following parts of the system may arise (SWJ+14): differences between drive wishes of the agent, the possible fulfillment of drive wishes in the simulation environment, emotions, and social rules. For instance, the agent's drives may aim to eat in a socially unaccepted situation. To resolve theses conflicts, psychoanalytic defence mechanisms are implemented, which filter and/or alter conflicted data. Examples of defence mechanisms are repression, denial, reaction formation, reversal of affect, displacement, idealization, and depreciation (SWJ+14). The general defense process is sketched in Figure 3.10. For conflict detection in the SiMA model

Super-Ego modules (F7 and F55 in Figure 3.8) are implemented, which scan the agent's drive wishes, emotions, perceptions, and social rules for conflicts.

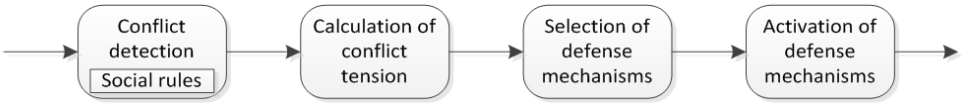

Figure 3.10: Functionalities of defence mechanisms (SWJ+14). After a conflict is detected and categorized, an appropriate defense mechanism is chosen that mediates between the conflicting sides (e.g. a drive and a norm).

In case of a conflict detection, first, the valuation of the conflicting components are taken as the conflict intensity (SWJ+14). Second, the agent chooses the defence mechanism to mediate the conflict unconsciously. Defence mechanisms, which have different degrees of maturity (from development) are selected based on the agent's current ego strength, represented by the amount of available neutralized intensity and a personality parameter.

Secondary Process Track

The main task of the SiMA secondary process is to make a decision based on the pluralistic information from the primary process. That is, in the primary process the functions process data under local conditions, e.g. every demand and norm is considered independently. The secondary process in turn, enables a global perspective on the different demands and affordances. This is done under the rules of the secondary process, i.e. the reality principle (how to avoid unpleasure and gain pleasure a la longue in the current and expected external world) and processing of hierarchical and causal data structures. In this regard, data structures are extended with word presentations, enabling logical thinking and verbal communication. Also, temporal and hierarchical associations may be used, making it possible to order things. At the beginning of the secondary process, activated stored images, which were independent in the primary process are formed into sequences called acts. Acts define events and the actions necessary to be taken to get from an event to another (DBD+15, pp. 93).

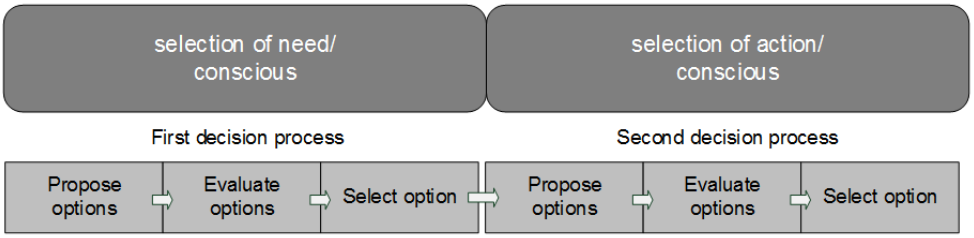

Figure 3.11: SiMA secondary process (SWK+15). See text for description.

In a general sense, decision making in the secondary process can be separated into two stages (SWK+15) (see Figure 3.11): Decision options are reduced, so that the remaining options have access to more system resources. First, all possible options have to be extracted from the processed data (drives, perception, memories). This is done in the 'Selection of Need track' in Figure 3.11 through the creation of possible goals ('Propose options' in Figure 3.11) from memorized activated acts or from perception. Drives define the desired external object, the preferred action and its importance. After a general effort analysis, all currently possible goals are evaluated regarding their possibility to fulfill current drives, emotion, and social rules ('Evaluate option' in Figure

3.11). Dependent on the available system resources (i.e., neutralized intensity) one or more possible goals are chosen for further processing ('Select option' in Figure

3.2.6.3 Information Representation in the SiMA model

The SiMA agent can be described as a memory-driven agent. An advantage of using psychoanalysis for the SiMA model is its provision of an abstract concept for memory structures, which is used for modeling the information representation that is defined in the scope of SiMA as follows:

"Information representation summarizes the structural composition of data that is received by the internal and external sensor system and the information management system" (Zei10, p. 9).

The central concept regarding information representation in psychoanalysis is a *memory trace*. In the SiMA project a memory trace is defined as "*a psychophysiological concept of representing memories in the psyche*" (Zei10, p. 49), (DFZB09, p. 424). A memory trace is a pattern for psychic data structures. Hence, memory traces form the base for all psychic data structures but are not themselves data structures (Zei10, p. 49). Incoming perceptions are matched against memory traces and activate them in case of a match. Data is then formed based on the incoming sensor data and activated memory traces. In case perceived information cannot be matched, new memory traces are constructed and stored. The concept of a memory trace implies that memory content is stored in an associated manner, with consideration of its co-occurrence, similarity and accessibility (Zei10, p. 49).

Psychic data structures are separated in thing presentations, which are processed in the primary process, and word presentations, which are processed in the secondary process. Their structure follows the rules of the respective process. For instance, thing presentations are not associated in structured form nor in any logical relation. The only associations used are co-occurrence and similarity.

Since the task of this thesis primarily considers primary process data structures, the remainder of this chapter will focus on them.

The psychoanalytically inspired technical concept of the data structures in SiMA distinguish between atomic data structures and composed data structures (Zei10, p. 50). Atomic data structures in the primary process are thing presentations and associations. Associations are weighted connections between data structures. Regarding object representation, a thing presentation (TP) evolves out of neuro-symbols and hence represents sensory information in a symbolic form. It is defined as "... the psychic representation of an object's sensorial characteristics in the form of acoustic, visual, olfactory, haptic, and gustatory modalities." (DFZB09, p. 426), (Zei10, p. 54). The sensor modality type, from which it originates, and an attribute value is the minimal definition of a TP. Associated TPs form the most important composed data structure of the primary process, a thing presentation mesh (TPM).

As it is assumed that every memory trace is created only once in memory (Zei10, p. 56), a TP may be associated to different TPMs. In its minimal form a TPM represents a physical object, which is described by TPs that are associated through attribute associations. In this regard class attributes are distinguished from instance attributes. The former are essential for the physical definition, the latter are individually different. Additionally to attribute associations with TPs, a TPM may be attributed by TPMs. This is the case if a physical object has distinguishable object parts, which are also represented as TPMs.

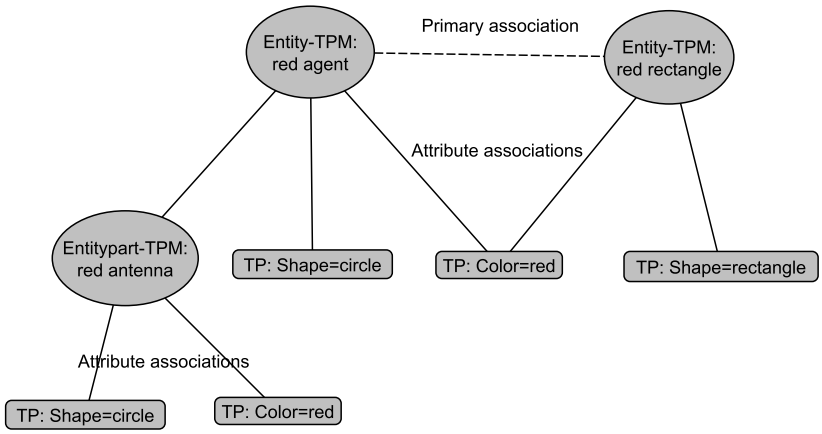

Figure 3.12: Simplified depiction of a thing presentation mesh in the SiMA model v38g.

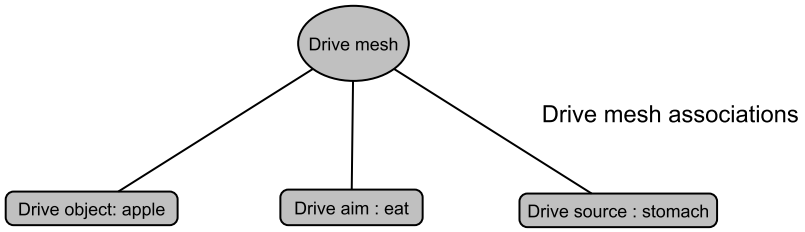

Figure 3.13: Simplified depiction of a drive mesh in the SiMA model v38g.

Different TPMs may be associated with each other through primary associations, representing similarity and temporal associations, which reflect co-occurrence of objects. To distinguish physical objects from other TPMs, they are called entities or entity-TPMs and TPMs that form representations of object parts are called entity-parts or entity-part-TPMs. A composition of entity-TPMs - representing a situation - is called an image-TPM.

Word presentations (WP) and their compositions, Word Presentation Meshes (WPM) are the equivalent to TP and TPM in the secondary process. Using these structures, symbol systems, e.g. language, is composed. Opposed to TPMs they may be associated in a hierarchical way. Therefore association predicates (e.g. 'is a') are used. TPMs that are used as WPMs in the secondary process may be associated as so-called acts for the sake of representing a sequence of events. An act is defined as "... a causally and logically traceable sequence of events which is assembled from images and motions and can contain other acts. It has one or more goals and may optionally include certain places (landmarks) relevant to the recognition of a situation. Goals and landmarks are likewise defined via images." (DBD+15, p. 96) And a goal is defined as "... a word presentation mesh which additionally transports contents from a drive representative. Alternatively, the goal can be associated with an image from perception, in which case it represents the possibility to satisfy a drive wish through the object represented by this image" (DBD+15, p. 97).

A drive mesh (DM) is another composed data structure that is processed in the primary process. It represent the psychoanalytic concept of a drive representation. As already mentioned, a drive is the psychic representation of a bodily need and comprises a drive source, a drive object and

a drive aim. Technically the components of a drive are associated with the drive mesh by drive mesh associations (see Figure 3.13). As already mentioned, the quota of affect reflects the drive tension in the drive source. Regarding further processing of the drive mesh the quota of affect represents the potential of drive satisfaction (i.e. reducing the drive tension in the drive source) by using the drive object in the drive aim's act. In case of a memorized DM the quota of affect reflects the amount of pleasure the associated drive object has brought in using the associated drive aim. In particular, it represents the degree of reducing the according bodily need.

3.3 Emotion

When searching for the building blocks of human decision-making, emotion can be regarded as an assumed foundational factor. On the one hand, foundational building blocks are generally valid for human decision making (individual or social). On the other hand, emotion as a mechanism probably evolved due to adaption to a social (i.e., dynamic) environment (e.g. (GM14, p. 1ff.), (SS13)) and is an essential part for establishing social behavior (Fri08). To understand the purpose of emotion for the sake of developing a computation model, before evaluating other computational models of emotion, a basic and broad interdisciplinary literature analysis is given. A special interest lies in how different approaches capture the concept of emotion. Based on that, an overview of computational models should give insight of how they use emotion, what their methodological context is, and where their innovation lies.

3.3.1 A First Grasp on the Concept of Emotion

Since ancient times, philosophers (e.g. Plato's 'Phaedrus') hypothesize that human behavior is determined by mechanisms that operate with different, even opposed, rules. Although different terms exist (e.g., Plato distinguishes affect, conation, and cognition), concepts to describe and explain the distinction between these mechanisms are commonly termed as rational and emotional. Especially since the enlightenment and the rise of science, human behavior is examined by rational means, in turn implying that reason is the only worthwhile means of the human mind to analyze, and even regarding emotions as a distortion of reasonable thinking. In this regard, emotions where seen as rationally not graspable, and stayed in the domain of the mysterious. Only recently science recognized the relevance to include models of emotions in their explanations. This is especially the case for decision-making. In this regard, even a turn in science can be observed: No decision-making is possible without emotions (Dam03). In computer science, at least since terming the concept of 'affective computing' by (Pic97), emotions incrementally became of interest for interactive applications.

The definition of emotion is highly influenced by folk psychology (concepts such as 'fear', 'joy' from everyday language) and introspection. This is also reflected in the disorientation of emotion research, where more than 150 theories in psychology and philosophy exist (Str03) (Sch05), with sometimes completely different understanding of emotions. The non-scientifically established term 'emotion' is hard to grasp in scientific terms. Since it is a sensitive and unconscious part of our everyday life, emotion is difficult to grasp - hence the pervasiveness of emotion and its folks psychology. Arguably, one reason why we even speak of something as emotion (and thus research them) is to have a concept for regular recurring states and response patterns in humans and to enable to capture and interpret other persons using this concept. But this typical folk psychology

approach is a descriptive and vague auxiliary concept. The pervasive usage of emotions as folk psychology terms is also mirrored in the usage of the plural, which implies that a single concept of emotion does not exist.

One approach for disciplines is to relate their theories to folk concepts from language ('joy' etc.), since they are dependent on probands' description of themselves in examining emotion. And since emotion is a concept difficult to grasp, folk concepts provide an available point of departure, in the sense of: emotions are what people say they are. However, a representative system in the mind that serves as the mechanism in which people think about emotion must be described using scientific dimensions, which again needs terminological and conceptual clarification.

Most approaches that tackle the definitional difficulties are conducted on a descriptive and phenomenological level. (Sch05) gives an example of aiming for an universal, invariant, and consensual working definition. A component process definition of emotion is given and demarcated from other 'affect states'. Emotion is defined as "... an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism." (Sch05, p. 697), with the subsystems serving as emotion components when they interact with other subsystems (which function independently when not included in emotion), as shown in Table 3.1.

Emotion function	**Organismic subsystem and major substrata**	**Emotion component**
Evaluation of objects and events	Information processing (CNS)	Cognitive component (appraisal)
System regulation	Support (CNS, NES, ANS)	Neurophysiological component (bodily symptoms)
Preparation and direction of action	Executive (CNS)	Motivational component (action tendencies)
Communication of reaction and behavioral intention	Action (SNS)	Motor expression component (facial and vocal expression)
Monitoring of internal state and organism–environment interaction	Monitor (CNS)	Subjective feeling component (emotional experience)

CNS = central nervous system; NES = neuro-endocrine system; ANS = autonomic nervous system; SNS = somatic nervous system.

Table 3.1: Relationships between organismic subsystems and the functions and components of emotion (Sch05, p. 698).

Hence, this concept of emotion consists of various components, feeling being only one of them. By comparing the demarcation problem of affective phenomena with that of demarcating language from other communication mechanisms, Scherer (Sch05) defines preferences, attitudes, affective dispositions, and interpersonal stances. He demarcates them from emotion by describing the intensity of design features that are derived from his emotion definition: event focus (concerning the external or internal emotion eliciting, event), appraisal (relating the event to one's concerns), response synchronization (synchronizing the subsystems for a response to the stimulus), duration (e.g. emotions last short and moods last long), intensity, behavioral impact, and rapidity of change. From this general definition, specific emotions may be described. Theoretically, there are as many different distinguishable emotions as there are distinguishable profiles of above fea-

ture descriptions. But which emotions should be distinguished and how to call them? Here, Scherer (Sch05) suggests to return to the folk concepts of emotion and define/map them using the component process description and their features. In the end, verbal distinctions of emotions are necessary and helpful. Hence, the decisive element in distinguish these emotions are appraisal dimensions, from which goal conductiveness (including valence) and coping potential (control/power) have the highest impact on emotion differentiation.

Scherer's (Sch05) definition is paradigmatic for psychological models. Emotion is regarded as one of many affective phenomena, with the key difference that it is elicited primarily from the external world and is only of short durance. The distinction between theses affective phenomena (mood, emotion etc.) seems functionally irrelevant, and serves only the connection to folk psychology, i.e., to relate this definition to empirical input from people. Even if functional explanations are integrated (with the appraisal aspect), the definition of emotion and the separation in different categories are done on a phenomenological level that is not grounded in a functional level.

Instead of a descriptive approach conventionally used in psychology, a computational approach is able to support a functional definition of emotion and supports understanding by forcing to sharpen the theories. On the other hand, the need of terminological clarification is especially obvious when following a computational approach. (RHD+13) suggest to systematize and classify the *assumptions* of emotion theories, formalize them in implementation-independent formal languages, and model emotions in cognitive agent architectures.

One possibility to get an overview of definition of emotion is to classify existing emotion theories. Although the decision of how to classify emotion theories also implicitly mirrors basic assumptions about them, it supports demonstrating the basic ideas in the field. Beside a distinction based on scientific discipline, different classification-systems are chosen. (Pan98, p. 43ff.) distinguishes three three general research strategies. The categorical approach assumes a set of basic discrete emotions, which are distinguished according to an "... objective analysis, behavioral expressions, human subjective experiences, established brain systems, or a combination of the above ..."(Pan98, p. 44). Opposed to that, the social-constructivist approach argues that no universal basic emotions exist, and that emotions are learned labels of bodily sensations and psychic experience. The componential approach represents a hybrid position that focuses on the appraisal processes triggering emotions, which are constructed from elementary visceral-autonomic sub-units. Another approach is to classify emotion theories according to the processing-principle, e.g., physiological or cognitive. That is, to consider emotion as a physiological or cognitive phenomenon. Taking the focused aspect and goal of the theory, one may distinguish design, semantic, and phenomenon-based models (BA08). Regarding the used defining components of emotions, theories are distinguished as discrete versus dimensional, communication-driven versus process-driven, horizontal versus vertical (RHD+13). (MGP10) distinguish emotion theories in regard to components that are considered as intrinsic to an emotion (e.g., cognition processes, somatic processes, behavioral tendencies and responses), the relationships between components (e.g. does cognition precede or follow somatic processes), and regarding their representation (e.g. is anger a linguistic fiction or a natural kind).

Such classifications only provide an overview of how emotion is currently approached in literature. To be able to relate the state of the art to this thesis' approach, in particular to evaluate appropriate emotion theories for the translation in the approached computational model, questions - derived from the introductory sketched criteria - are specified.

3.3.2 Analyzing Emotion Theories for a Computational Approach

To evaluate against the criteria of an embodied, generative-functional, and holistic-unified computational approach, the following questions are important: (1) What does the term emotion describe and what is the defining composition of emotions? Answering this question helps to represent emotions scientifically. (2) How do emotions arise? (3) What is the claimed purpose of emotions? These intertwined questions are tackled in the remaining sections, where paradigmatic answers to these questions are given.

3.3.2.1 Explanandum and Components of Emotions

A first hint for what term emotion is used to describe and explain (i.e., the explanadum) is given by its etymology[18], i.e., describing a mechanism that moves or drives a person. As introductory mentioned, in most emotion theories folk concepts are used as a point of departure.

Behavioral approaches aim to map folk concepts (e.g., fear) to *external processes*. That is, they use behavior and expressions as defining components of emotion. Such theories claim that the concept of emotion is immanently behavioral and have to be described primarily as a phenomenological category (emotions as folk concepts are phenomenological), because they cannot be reduced to cognitive aspects (SS13, p. 25). In this case a specific kind of expression or behavior is called emotional, since it is assumed that the underlying mechanism does not follow reason, but other principles that are best describable by their behavior. In this regard emotions provide a concept to explain a specific type of behavior. However, such approaches can have the same faults as behavioristic approaches, limiting their research topic on objectively observable (external) behavior.

Paradigmatic concepts for behavioral approaches are mostly based on Darwin (Dar98), who researched how behavior expresses emotion. Darwin describes three principles to explain facial expressions and gestures (SS13, p. 85) (Dar98, p. 33ff.)): "The principle of serviceable associated Habits" states that movements that serves satisfactions of needs or are otherwise helpful become a habit; "The principle of Antithesis" says that antagonistic behavior is expression of antagonistic states; "The principle of actions due to the constitution of the Nervous System, independently from the first of the Will, and independently to a certain extent of Habit" explains side effects such as tremor. A contemporary theory of this approach is the 'facial feedback hypotheses' (McI96), which argues that facial expressions are necessary components of emotions and correspond to emotions. Experiments, such as the impact on an person's state when holding a pen in her mouth (SMS88), provide evidence for that claim. A taxonomy for a facial-based emotion theory is given by (Ekm99) by defining distinctive 'basic emotions', which are identifiable via their facial expression and culturally-independent.

Generative approaches (process models) aim to map the folk concepts of emotion to the underlying *inner processes* that generate that behavior or expression. For instance, neuroscience searches for the neural substrate and systems of emotion. Some psychologist try to abstract from the neural substrate and describe those underlying systems on a psychological level. Here the focus lies on mental states as defining aspect of emotion. Disciplines such as Cognitive Neuroscience and Neuropsychoanalysis aim to bridge classical neuroscientific and psychological

[18] until the 19th century in German the term Gemütsbewegung (spiritual movement) was used instead of Emotion, which emphasizes the etymological meaning of the word

approaches. Overall, such theories argue that emotions are best described as mental mechanisms, since that is what they *primarily* are. However, it may be problematic to take folk concepts as a point of departure, as usual done.

Panksepp was one of the first that looked for the neural mechanism of emotions in a systematic way (Pan98), coining the term 'affective neuroscience'. Emotions are defined to "... arise from executive circuits of the brain that simultaneously synchronize a large number of mental and bodily functions in response to major life-challenging situations ..." (Pan98, p. 123). Their purpose is to constitute coherent action plans, supported by adaptive physiological changes. Overall, Panksepp considers emotion as an umbrella concept for affective, cognitive, behavioral, expressive, and physiological changes (Pan98, p. 32). To define emotions, examining the core substrate of emotion, i.e., neural systems, allows for deeper explanations - opposed to using "peripheral" (Pan98, p. 14) factors, such as behavior and bodily expressions. Nevertheless, as a reference point Panksepp uses "... discrete types of instinctual behavior that can be conditioned and the neural systems from which they arise" (Pan98, p. 13). In fact the conclusion is that the underlying emotional systems that Panksepp found "... appear to have common characteristics with everyday emotional concepts" (Pan98, p. 15). Theses statements are the result of decades of animal experimentation, with the key insight that all mammals share a set of seven distinguishable innate emotional systems (Pan98, p. 51), from which the major emotional operating systems are SEEKING, FEAR, RAGE, and PANIC[19] (Pan98, p. 53). As sometimes done in neuroscience (cf. the Triune Brain Model (Mac90) (Wie10, pp. 9 ff.)), these systems are assumed to be based on evolutionary older reflexive instinctive abilities (Pan98, p. 47). The mixture of these basic emotions and their interaction with higher brain functions generate complex emotions (Pan98, p. 48).

However, Panksepp mentions that basic emotions are not able to capture all affective feelings, colloquially called hunger, thirst, tiredness, illness, surprise, disgust (Pan98, p. 49). Another conceptualizing system must be found for them. And even with basic emotions a taxonomy must be kept open for new research findings. However, even if Panksepp uses discrete neural systems to represent emotions, he states that they share a common dimensional representation (Pan98, p. 46).

The question of scientific representation of emotions, in particular how they may be represented formally is of high interest for a computational approach. Are they represented implicitly or explicitly? Is the function or structure represented? As implied in the above sections, typical approaches of representations are discrete or dimensional. This design-decision implicitly defines the intrinsic primitives of emotion that a theory states. Besides, the presented neurological approach to distinguish basic emotions, multiple psychological separations exist with the claim of a universally valid minimum set of emotions. These basic emotions are supposed to be the primitive building blocks of emotional states.

Opposed to a discrete distinction of emotions, a dimensional approach to emotion doubt that the full spectrum and complexity of emotional states are representable by a discrete conceptualization. The founding father of this idea is Wilhelm Wundt (Wun22), who proposes to define an emotional space that is constructed by three orthogonal dimensions, namely pleasure-unpleasure (Lust/Unlust), excitement-inhibition (Erregung-Beruhigung), tension-relaxation (Spannung-Lösung) (BA08, p. 23). Hence, an emotional state is represented as (sub)space in this three-dimensional space. As opposed to the emotion theories described so far in this thesis, Wundt's idea is the first one that does not use folk concepts or external features as point of departure but internal

[19] Panksepp uses capital letters for the underlying neural emotion systems to distinguish them from verbal labels.

experienced features of an emotional state. Such an introspective approach (the other extreme are behavoristic approaches, see above) has been criticized in his lifetime, as it has been today. However, in the last decades a vast amount of empirical research was conducted about the appropriateness of each dimension. The idea that two dimensions, pleasure-displeasure and activation-deactivation, are enough to represent all internal states dominated for some time (Rus03). After demonstrating (using surveys) that different reported emotional states are difficult to distinguish with two dimensions, e.g., anger and anxiety, a third dimension (dominance-submissivness) has been introduced (MR74). Most current dimensional approaches consensually use these three dimensions, which can be described as valence, activation, and potency (the person's assessment of potency in dealing with an event). A prominent example is the PAD model, which describes the three dimensions as Pleasure, Arousal, and Dominance (Meh96). These dimensions are able to represent emotional states of different durations. Assumedly, a person is in exactly one emotional state at any moment (Rus03). The basis of these different emotional state can be described as 'core affect', which is a experiential state of valence that is expressed as different emotional states (e.g., mood, emotional event) (Rus03). (RHD+13) recognizes the lack of intentionality in the concept of core-affect as the main difference to most emotional theories. This is also reflected in de-emphasizing the term emotion and use it to describe the cognitive usage of core affect (MGP10).

3.3.2.2 Purpose, Functionality, and Function of Emotion

With the premise that emotion serves a functional mechanism with possible non-functional effects, an analysis of this aspect is central for the translation of an emotion theory into a computational model. The purpose of emotion is described quite broadly in the literature. Most theories regard emotion as serving adaptivity. In a general sense, emotion serves the preparation to act or respond to demands, and orients a person to motivational important cues in the environment (Can98). Emotion is also regarded abstractly as an information-providing mechanism (CGG01). (Pic97, p. 298) describes emotion as a representation of the internal state. The 'affect-as-information' hypothesis regards affect as embodied information about value and importance (CGG01). In this regard emotion governs the information flow by informing about the importance of cognitive states or objects, serving as "cognitive bookmarks" (Meg14). The concrete purpose in information processing is classically described as interrupting mechanisms focusing the mind's attention to relevant events (Sim96), which can be regarded as the beginning of a "cognitive theory of emotion" (Wen05). Another established general purpose of emotion is its integrative function. In this regard the dynamic character of emotions is emphasized, where emotion integrates "... causally related processes of several subsystems: the physiological, the cognitive-evaluative, the communicative-expressive, and the subjective experience subsystem ..." (Can98, p. 2). This view of emotion would consider it as a meta-function. In this regard also neuro-scientific accounts argue for emotions as an integrator of different functionality (Pes08). Pessoa argues against the separation of cognitive and affective brain areas and demonstrates that complex cognitive-emotional behavior is based in dynamic interactions between brain networks. In particular, between brain areas with a high connectivity - so called hubs - that are central for the regulation of the integration and distribution of information between brain regions. Pessoa demonstrates her hypothesis with the amygdala. But given the integrative function of emotion, i.e., integrating information coming from different (internal and external) sources, Pessoa argues against a conventional mapping between (static) brain area and function. Instead she suggests a mapping of (dynamic) connectivity patterns to functions. Another view on this proposal is to avoid using descriptive concepts (such

as emotions) for functional modules (such as the dynamic function of the amygdala). Others argue that "... the components of emotional syndromes are in fact only loosely associated" (Bar06)(RHD+13, p. 252) and the seemingly integrative aspect of emotion is the result of its single functional effects. In this regard, different functions are proposed: (RHD+13) summarizes them using three categories. (1) The informational function considers that emotion provides adaptively important information (e.g., about value) to other systems (although this would again emphasize the integrative function of emotion). (2) The attentional function regards emotion as guiding the mind's processing on relevant events. (3) The motivational function can be regarded as a goal-generation functionality and uses functions from (1) and (2). Another summary of emotion's purpose is adaptation, social communication, and subjective experience (Can98).

Such categorizations emphasize the purpose and functionality of emotion, but not its (information-processing) function. The literature is in consensus about the adaptive purpose of emotion, and - as sketched above - proposes different functionality that is able to account for that adaptivity. But what does the literature say about the underlying information-processing functions? What are the processing principles? This question can be tackled by different aspects: low-level (i.e., body near) processing aspects (e.g. motivations), high-level aspects (e.g., decision-making, higher cognition), and cross-level aspects (e.g., social aspect, learning).

Motivation

A recurring mentioned functionality in literature is the motivational aspect of emotion, which is also reflected in the term's etymology. These two concepts, motivation and emotion, should not be considered separately, e.g., by considering emotion as output of motivation potential (IBH15, p. 3). Generally, motivation can be regarded as the function of goal generation and prioritization. Almost all theories of motivation are variants of the belief-desire paradigm (RHD+13, p. 252). The central idea here is that actions result from beliefs about one's state and desires for changing that state. Hence,informational and motivational factors are needed. In this framework the concept of emotion is not required, and not surprisingly most of motivational theories do not consider this concept. This goes in line with insights from artificial life simulations (PP10), which demonstrate that emotions are only a supportive mechanism for motivation. However, such simulations also show that, even if the concept of emotion is not a key factor of motivations, it complements and supports desires - a relation which is evolutionary comprehensible. A similar relation between motivation and emotion is implied in Reisenzein2013, where the evolutionary function of emotion is described as improving adaptive action generation beyond the possibilities of beliefs and desires.

The role of emotion, of course, is dependent on the usage of the concept. The motivational aspect of emotions is also used in some cognitive architectures (see Chapter 3.2.3 3.2.4). And here, opposed to (PP10), artificial life simulations with Psi agents showed the relevance of emotions for the survival in virtual worlds, although the Psi architecture is also primarily a theory of motivation (similar to the belief-desire paradigm), and considers emotion only implicitly. Hence, the relevance of emotion in such systems is implicitly dependent on their given relevance by the designer. For instance, (FR06) considers motivation, valuation, and emotion as three sides of the same coin: "Motivation primes action; value serves to choose between motivations; emotions provide a common currency for valuation and implement motivation", emphasizing that all motivations and emotions must be in service of a prime motivation for the sake of embodiment and grounding.

Decision Making

Classical, economically inspired, decision theories regard the decision-making process abstractly as maximizing an person's expected utility. Instead of only using the absolute final asset to choose

the best decision (cf. consequentialism), newer approaches consider the probability that the evaluated outcome will occur and the relative loss and gain for a person (LL03, p. 626). Nevertheless such theories typically are concerned with task models, without considering the underlying mechanisms of decision making. With emotions by and large ignored until the end of the twentieth century, recent decision theories emphasize their importance (LL03), without neglecting their potential for biased judgment. (LL03) summarizes two ways of how emotions enter into decision making and separates expected and immediate emotions in this regard. Expected emotions correspond to the cognitive assessment of possible outcomes of an decision, and can be regarded as the result of a reasoning process about a decision's utility. Most importantly, they are able to be felt, hence the provided information is purely cognitive. Immediate emotions (colloquially called 'gut feelings', which are mostly neglected in decision theories (LL03)), in turn, are the felt state during the decision making process. Most importantly, immediate emotions' impact consist of (1) action tendencies immanent to emotions, which can be regarded as hard-wired reactions (e.g. fleeing), and (2) mood influences, in particular the tendency of optimistic decision in a good mood and pessimistic decisions in a bad mood (LL03, p. 628). Two factors determine immediate emotions: Incidental influences, i.e. factors that are not related to the decision at hand, and anticipatory influences, which differ to expected emotions most importantly in neglecting probabilities of an outcome.

Other approaches that focus on the evaluative functionality of emotion, go beyond cost-benefit calculations, and consider various so-called appraisal variables in the decision making process (see below).

Higher Control

Instead of separating cognitive and emotional processes and regarding emotion only as an influence of cognition, (IBH15) argues the other way around: cognitive control can be regarded as an emotional process. In particular, the function of emotion is identified as recruiting cognitive control, which is defined as "mental processes that allow behavior to vary adaptively depending on current goals" (IBH15, p. 1), e.g., enabling long-term goals. that "cognitive control is initiated when goal conflicts arouse negative affect" (IBH15, p. 2), which focus and motivates goal-directed behavior to resolve the conflict (being a disagreement of different mental representations, processes, or behavior).

With the function of recruiting attention and mobilizing for action, emotion is an appropriate means for the functionality of recruiting control. However central emotion is, (IBH15) concludes that although it is necessary, emotion is not sufficient for cognitive control.

Appraisal

Opposed to the view sketched above ((IBH15)), appraisal theories regard emotion as the consequence of cognition. Following the common view that emotion is an reaction to an event, such theories (e.g., (OCC90)) focus on the processes that generates emotional reactions. This process is called appraisal, since it describes the subjective assessment of an event, action, or object, and is supposed to be cognitive. Appraisal theories are predominant in psychological approaches to emotion, since - given the emphasis on cognition in psychology - it claims to bridge cognition and emotion. However, it has to be seen as a purely cognitive approach to emotion, describing emotions solely from a cognitive view.

Others emphasize that emotions have to be seen from different aspects, containing different components. Accepting appraisal as one source of emotion, other approaches (e.g., (Iza93)) argue that non-cognitive elicitors have to be taken into account as well. Overall, one may criticize

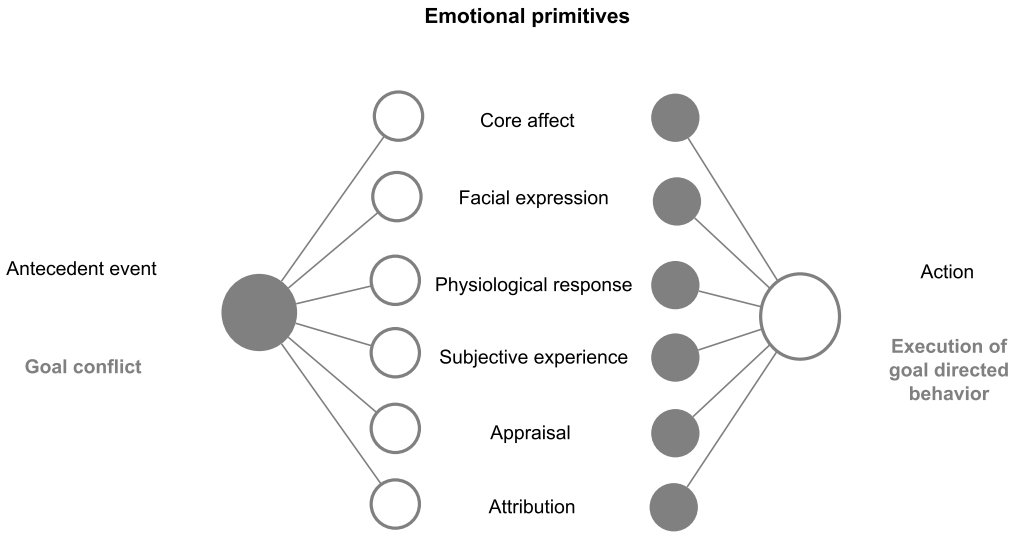

Figure 3.14: "Emotional foundations of control. When decomposed, emotional episodes break down to (i) an antecedent event that (ii) produces changes in a cascade of different emotional primitives, which then (iii) motivate the execution of goal-directed behavior. When cognitive control is similarly decomposed, it becomes apparent that it is constituted by the same types of elements. Specifically, cognitive control is constituted by (i) an antecedent event (goal conflict) that (ii) triggers a host of emotional primitives (including changes in affect, facial expressions, underlying physiology, subjective experience, appraisals, and attributions) that (iii) motivate refocusing on goal-directed behavior (recruitment of control)." (IBH15, p. 2)

the appraisal approach as 'overintellectualizing' emotion, neglecting embodiment and grounding emotion.

As their name imply, appraisal theories are concerned with describing the cognitive process of appraising primarily external events and situations against a person's beliefs, desires and intentions. This appraisal in turn triggers other cognitive responses, in particular coping strategies (e.g., planning, procrastination). Classically such theories focus on (1) defining a sufficient set of appraisal variables, and (2) try to map the processing of these variables into common emotion labels. Hence, appraisal theories are not so much concerned with a categorical model of emotion (i.e., emotion is no explicit concept), but claim to aim for a (cognitive) process model. Following evaluation criteria can be considered a consensus for the definition of appropriate appraisal variables (Ran09, p. 83): (1) relevance, goal significance, focus; (2) standards compliance, blameworthiness; (3) intrinsic pleasantness, valence, appealingness; (4) novelty, unexpectedness, suddenness, familiarity; (5) responsibility; (6) coping potential. However, the concrete process of the applying these criteria in evaluations, i.e. a concrete process model of them, is not described in psychological sources. These details are accounted for in computational models of emotion, which are mostly based on a form of appraisal theory, since it is appropriate for symbolic AI approaches. The relation between appraisal theories and AI is best demonstrated by the argument that it enables reasoning about emotion (OCC90). In this regard further details are given in the section concerned with computational models of emotion (see Chapter 3.4).

Empathy

Emotion and its usage in empathy is regarded as an essential part for establishing social behavior

(Fri08) and probably originally evolved due to adaption to a social (i.e., dynamic) environment (e.g. (GM14, p. 1ff.), (SS13)).

Emotion is the basis of those mechanisms that enabled the cumulative cultural evolution of humans, which can be regarded as a decisive turn in human history (Tom08, p. X). The cultural origin of human cognition (Tom99) is facilitated by the cooperative capabilities that lead to teaching, imitative learning, and the creation/recognition of a shared intentionality and attention (Tom08, p. XI ff.)[20]. On top of theses capabilities social institutions evolved, which are - beside cumulative artifacts - the second characteristic of human culture (Tom08, p. XII). Most importantly, these cooperative capabilities are innate and not a result of acculturation, which is shown in experiments with toddlers, who behave altruistically in helping others without parental intervention (Tom08, p. 11). Beginning in the age of three years, children recognize the need for selective usage of their altruism, and acculturate concepts of reciprocity and reputation (Tom08, p. 28).

These cooperative capabilities require a mechanism for mindreading. (Gol06, p. 13) distinguishes three approaches to explain mindreading mechanisms: theory-theory, modular-nativist theory, and simulation theory. Theory-theory follows a folk psychology approach in assuming that humans use naive psychological theories in reasoning about others beliefs, desires, and intentions. Modular theory assumes an innate specific theory of mind module (ToMM) [21] that mentalizes about other's mind. These approaches stem out of autism research, with the claim that autistic persons lack such module, but are still able for reasoning about their own beliefs, desires, and intentions. Baron-Cohen popularizes this idea and specifies four modules (Gol06, p. 15) (BC95): intentionality detector, eye-direction detector, shared-attention mechanism, ToMM representing epistemic mental states, such as pretending, thinking, knowing, believing, imagining, dreaming, and guessing." (Gol06, p. 16). Simulation theory assumes that a person uses the same mechanisms to emulate others mental state. The mind is assumed to do that by resembling the processes of the other mind (Gol06, p. 32), resulting in a final state that is isomorphic to the other mind. In this regard, after replicating the others mental state, the mindreader projects her own mental state to the other (Gol06, p. 40). Different ways of conducting such simulations are assumed, of which all assume emotion recognition and experience of their emotion as the basis (Gol06, pp. 113 ff.). The generate-and-test approach, where a person simulates others' assumed emotion, produce an according facial expression and test it against that of the other(Gol06, p. 125). The reverse simulation approach starts with imitating the other's facial expression and experience the resulting emotion and compare it with the other person's emotion (Gol06, p. 126)[22]. A variation of this approach assumes that an activation of an somatosensory representation of the other's facial expression (without the necessity to display these expressions) suffices to experience the other's emotion, which could be facilitated by an as-if-loop in the brain (Dam94). The unmediated resonance (mirroring) approach assumes that the perception of the other's face directly - i.e., without mediation by other processes, activates the neural substrate of the according emotion, resulting in a mental state that resonates the other's (Gol06, p. 127).

[20]Tomasello hypothesizes that following factors are characteristic for human cognition (compared to other animals) (Tom99, p. 22): (1) Identification with others (enabling recognizing others as intentional beeings as oneself). (2) This enables cultural learning and the cumulation of cultural artefacts. (3) Adaption on this cumulative culture.

[21]Theory of Mind (ToM) refers to the fundamental mindreading capacity, but does not say anything about how it is implemented, which is stated by the Theory of Mind Module (ToMM) (Gol06, p. 112).

[22]The difference between these two approaches resembles the century old question: Is emotion a representation of the body or the mind. That is, do we sweat because we are afraid, or are we afraid, because we sweat - hence sweating is not caused by cognition, but e.g. by perceiving an dangerous situation (cf. (Jam84)).

Unmediated resonance gets its evidence from the existence of mirror neurons. The original neuroscientific findings show how mirror neurons join representations of perceptions and actions by directly translating perceptive input into motor representation (Key13, p. 1, 81). This demonstrates the causal connection between perception and action in the brain (Key13, p. 67). Originally observed in the monkey premotor cortex as neurons that "discharge both when the monkey performs an action and when he observes a similar action made by another monkey or by the experimenter" (RFGF96, p. 131) and provide the basis for understanding others' action, the same principle is found in humans (Key13, p. 53). Furthermore, the same mechanism is assumed to account for intention recognition of others by activating others' behavior into (Key13, p. 42). Similar to mirroring perceptions of actions, the existence of mirror neurons in brain areas of facial expression and emotion (e.g. insula) are observed (Key13, p. 118). In this regard, mirror neurons in these areas account for synchronizing emotional state between humans, by translating facial expressions into emotions. These findings emphasize the need to experience other's assumed emotion by ourselves to be able to connect to their mental state. Hence, mirroring actions (see above) and mirroring emotions are assumed to be the mechanisms for empathy, with the premise of having experience (i.e., memories) of similar actions and emotions (including their situative context) (Key13, p. 137). Although the concept of mirror neurons provide important insights, the conclusions of their observation is questioned (Hic09) and put into context (KL13). The claim of being the basis of all social cognition or even of all mentalizing is also questioned, e.g. in (Gol06, p. 210).

After reviewing the assumed relevance of emotion and empathy for mindreading and hence human social existence, the ground is set for a general definition of empathy, independent from its implementing mechanism. Baron-Cohen defines empathy as "... our ability to identify what someone else is thinking or feeling, and to respond to their thoughts and feelings with an appropriate emotion" (BC12, p.12). Another widespread definition is given in (VS06, 435): "There is empathy if: (i) one is in an affective state; (ii) this state is isomorphic to another person's state; (iii) this state is elicited by the observation or imagination of another person's affective state; (iv) one knows that the other person is the source of one's own affective state". Similarly to the distinction between automatic unconscious) and controlled (conscious) mental processes, different levels of empathy are distinguished, such as affective and cognitive or low and high-level empathy (Gol06, p. 207). Other demarcations include concepts such as emotion contagion, sympathy etc. exist. Opposed to that (Waa12, p. 100) introduces empathy as an umbrella term, connecting automatic and deliberative empathy, sympathy, emotional contagion, perspective taking. Therefore de Waal proposes a layered model, with empathy encompassing all levels, depicted as a "russian doll model of empathy and imitation" (Waa12, p. 101) (see Fig. 3.15).

3.3.2.3 Damasio's Theory of Emotion: Towards a Holistic-Integrative Model

Damasio provides a theory of emotion that aims to integrate body and mind, and approaches the topic holistically. Instead of focusing on either cognitive or non-cognitive aspects, Damasio sketches a broad picture of emotion (e.g. (Dam03, p. 40ff.)). On the one hand, he emphasizes the role of emotion in decision making, on the other hand he grounds emotion in the body. Additionally he emphasizes its homeostatic purpose in life regulation and provides a framework that connects basic mechanisms of life regulation, motivation, emotion, and feeling.

Extensive neuro-scientific findings based on the so-called somatic-marker hypothesis showed the role of emotions in decision making under uncertainty (NSB06). Here emotion are seen as bodily

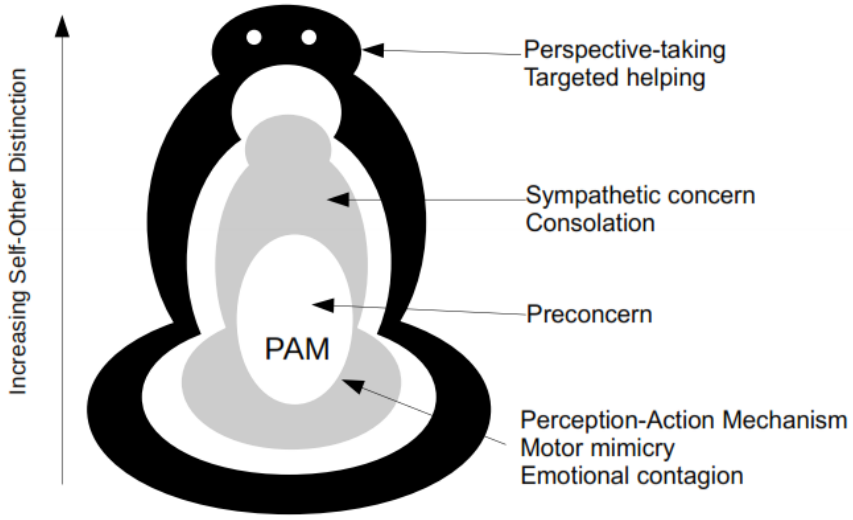

Figure 3.15: Russian doll model of empathy and imitation (Waa12, p. 101). The core of empathic mechanisms is the perception-action mechanism, with every additional outer layer building on the inner layers. The complexity of empathy grows with increasing self-other distinction in (human) development.

states, elicited during decision-making, marking the expected (dis)advantages of possible options. Damasio observed that patients with lack of emotional capabilities (regarding their behavior) are unable to learn form bad decisions due to a damage in a certain brain area (vmPFC: ventromedial prefrontal cortex), although their cognitive abilities seemed normal (Dam94). This lead to the conclusion that non-emotional decision making results in unfavorable, irrational, outcomes. Hence, emotion is a necessary guide for *reasonable* decision making. This hypothesis was most famously reinforced by the Iowa Gambling Experiment (BD05), where patient with vmPFC-damage were not able to learn from losses when choosing disadvantageous decks of cards. In other persons, where somatic markers are functional, these bodily states are linked to mental representations, so-called mental images, that elicit them (NSB06).

Damasios's theory of emotion focuses on the evaluative function of emotion and its embodiment. This can be regarded as unifying separated approaches of bodily and cognitive emotion theories. Emotion is a bodily phenomenon, elicited by cognition (NSB06). Even more, the adaptive function of emotion only works by translating the evaluative result of the psyche into an a bodily state and (by that) bringing the current state, and hence the decision to consciousness. And here Damasio also locates the separation of emotion and feeling, with the later being the consciously perceived bodily state elicited by the evaluative process (Dam03, p. 44ff.). However, similar to Panksepp, feelings can be more than perceived emotions. This separation is mirrored in attributing emotion as objective and feeling as subjective. That is, emotion is defined as the perception (which is always an interpretation) of one's (bodily) state by other persons, and feelings the perception of one's (bodily) state by oneself. Such interpretation emphasized the social core of emotion.

This mechanism of emotion can be regarded as an operationalisation of a decision in the body, in the world. Or in other terms: it can be regarded as *embodied cognition*. The emphasis of the body as the medium for emotion to work reflects the evolutionary necessity: Evolved mechanisms have to fit in to given ground. And the common denominator of this given ground for mechanism that we call motivational, emotional, cognitive etc. is the body. These mechanisms - in an

evolutionary sense - serves a common purpose, namely automatic life regulation (Dam03, p. 37ff.), and hence need to be integrated. As long as, in the process of this integration, the usage of established mechanisms (e.g., the bodily path in life regulation) prove adaptive, these mechanisms are build on each other. Similar to cybernetic homeostatic mechanisms, regulation is triggered by internal or external changes. Hence, Damasio regards the human as a homeostatic system with emotions as one of the evolved automatic mechanisms of life regulation. To transport how the evolved mechanisms relate to each other and are integrated in a holistic system, Damasio uses the metaphors of a tree and nesting (Dam03, p. 40ff.). In summary, the lower levels of homeostatic regulation follow simple reactive: approach-avoidance, the higher levels follow competitive or cooperative reactions.

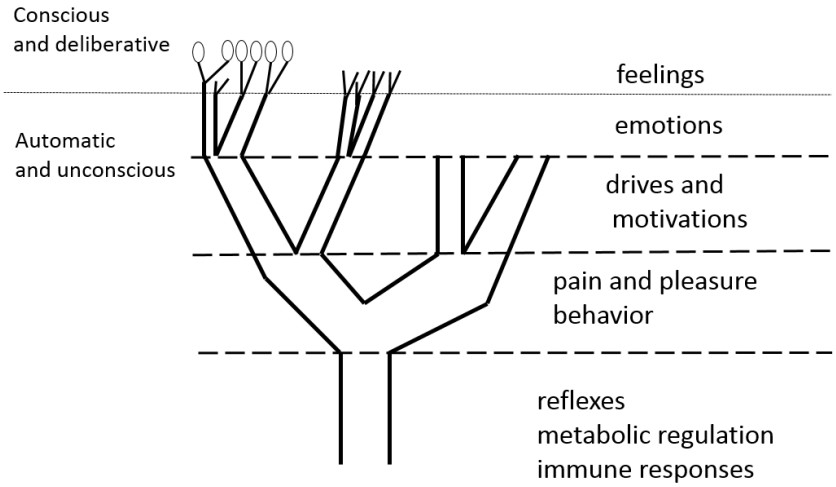

Figure 3.16: A tree as a metaphor for nested homeostatic mechanisms of life regulation, after (Dam03, p. 37). Later evolved mechanisms of life regulation include, use, and refine older mechanisms for the sake of increasing adaptivity.

The first level is considered with chemical homeostasis. The second level is the first psychic level and is concerned with general approach-avoidance behavior based on providing a system of pleasure as reward and unpleasure as punishment. However, on this level the feeling of pleasure and unpleasure is not the reason for the approach-avoidance behavior, as for instance in conditioning. The third level is represented by the drive concept, which reflects self-preservation, sexual desires, and curiosity (Dam03, p. 45). Basic emotions represent the fourth level of regulation. All these homeostatic mechanism are innately active. However, humans learn how to apply them appropriately in their environment (Dam03, p. 45).

Most importantly, processes of lower levels are *partly* included and used in higher-level processes (Dam03, p. 49). However, the higher levels do not include the lower levels completely. Nature have not evolved that systematically, but rather can be compared as a patchwork (Dam03, p. 49). So, the relation of the levels is not linear. That is why Damasio rather choose the metaphor of an irregular grown tree instead of a systematically developed building. In a technical sense, we would say that the interfaces between the levels do not follow a common pattern, but have to be specified between every level separately, and that different levels may have interfaces with each other (e.g., feeling and drives). The partly nesting of hoemostatic mechanisms also emphasizes that higher control is in the service of lower mechanisms and serves the same homeostatic regulation with different means.

The consideration of different evolved brain regions and the integration of their functionality

is a common principle in neuroscience. Such hierarchical models of the brain can be found in basic neuroscientific models, beginning with the separation of reptilian, paleomammalian (limbic system), and neomammalian brain in the triune brain model (Mac90) (Wie10, p. 24).

3.4 Computational Models of Emotion

After analyzing emotion theories, we take a look on why and how they are used in computational approaches. What is the general approach in using concepts of emotion and why is it needed in artificial systems? After that paradigmatic computational model of emotion are presented and finally discussed regarding this thesis' criteria.

3.4.1 For Science and Engineering

The reasons why computer science is interested in emotions can be distinguished in two categories, which resembles the traditional (artificial) separation between AI (using computational means to build intelligent systems) and Cognitive Sciences (using computational means to understand the mind). The AI approach aims to harness the functionality of emotions for an artificial system. And even if the requirement for such functionality for intelligent agents is argued (e.g. (Pic97), (Can98)), such approaches emphasize that endowing agents with emotion is not an end in itself (Can98). An agent's functionality has to be adapted on its environment and on its task. Otherwise (Can98) argues that such systems would be overdesigned, and unnecessarily anthromorphistic. In particular, the emotional capabilities should be adapted on the complexity of the system-environment interaction. Still, the question remains, which criteria an artificial system has to fulfill to be called emotional. One general answer is: whenever the functions that emotion serve (see above) are replicated.

Computational approaches as introduced in (Pic97) usually harness emotion for generating behavior that user would regard as emotional, with a focus on expressive display. The main purpose of emotions in such systems is to create believability, a term coined by (Bat94) to emphasize the need for convincing, life-like agents, from the perspective of a human user. Prominent application fields are in HCI (Human-Computer Interaction), conversational agents, entertainment, serious gaming. However, although such systems may express emotion realistically in their constrained application field, they do not represent them realistically, i.e., the do not *have* emotions (PP10).

Additionally, before we are able to identify the factors and aspects appropriate for single artificial systems and tasks, first we have to analyze functions, impacts, and capabilities of emotions, which is still an open research field. So using an computational approach to support the understanding of emotion, is a premise for applying them in artificial systems. Such scientific approach of using computational methods to tackle research questions form psychology can be regarded as synthetic psychology (Bra84) and is able to demonstrate hidden assumptions and complexities of emotion theories (MGP10), explication of theories, resolving inconsistencies and ambiguities, and adding missing assumptions (RHD+13). Generally, computational methods enable feasibility studies, and support model validation and exploration. Additionally, given the dominance of the computational paradigm in Cognitive Science, computational concepts support theory development.

However, to realize the potential of computational methods, interdisciplinary collaboration must be intensified, as well as the intra-disciplinary systematization (RHD+13). Possible proposals to

achieve this are (RHD+13): (1) to systematize assumptions of emotion theories; (2) to use set theory or agent logics (such as BDI) to formalize emotion theories in an implementation-independent representation; (3) to use general-purpose (cognitive, emotion, agent) architectures to model emotions. These proposals serve modularization and standardization. Otherwise computational models would just resemble the problems of the scattered and non-cumulative emotion theories in psychology. Using a common conceptual framework would also guide knowledge translation from psychology to computer science and possible identify shared core assumptions and building blocks (RHD+13). However, some aspects must be considered: Using an implementation-independent formal representation implies the feasibility of an analytical approach to validate emotional theories. But given insights from complexity theory, we know that complex models cannot be 'solved' analytically, but must be simulated (see Chapter 3.1.2). The usage of cognitive architectures to provide a basic infrastructure of domain-independent computational processes and structures for emotion modeling is an interesting approach in this regard (RHD+13). Nevertheless, such architectures have not considered emotions in their basic concept (see Chapter 3.2). Hence, in most cases (if the architecture is not too general) a cognitive architecture cannot be extended but must be merged with an emotion model. This mirrors critic on established cognitive architectures (see chapter 3.2).

Generally, one can observe that computational models of emotion again just resemble the different approaches in psychology, and favorably use theories applicable for implementation, instead of using computational concepts and methodology for basic clarifications in the emotion theories. This is also reflected in their design approaches.

3.4.2 EMA

EMA (EMotion and Adaptation) is a one of the most-referred appraisal-based models of emotion. It follows the standard approach of appraisal theories in regarding emotions arising from an agent's interpretation (i.e., appraisal) - mediated by cognitive processes and describable by appraisal variables - of its relationship to the current environment (MG09). Specific emotions then are represented by configurations of this appraisal process. But different to other appraisal theories who largely focus on describing the structure of appraisal (e.g., which appraisal variables to choose and how to represent them), EMA focuses on the dynamic aspect of emotion, i.e., the appraisal process itself (MG09). Also, different from other models, such as Scherer's appraisal model (Sch05), appraisal is regarded as a parallel, not sequential, processes. In contrast to multi-process models, appraisal is regarded as a fast, single-level process that can exibly utilize the result of perceptual and inferential cognitive processes, which may be fast and automatic or slow and deliberate (MG09). Hence, appraisal and inference are regarded as distinct processes that operate over a shared representation of an agent's relationship to its environment.

EMA uses the so-called causal interpretation - consisting of beliefs, desires, intentions, plans and probabilities - to represent an agent's current view of its relation to the environment (MG14). The appraisal process, then, is defined as "a set of continuously active feature detectors that map features of the causal interpretation into appraisal variables" (MG14, p. 59). Appraisal variables are relevance, perspective, desirability, likelihood, expectedness, causal attribution, controllability, and changeability. To calculate some of these variables, EMA uses decision-theoretic concepts of utility and probability. Relevance is represented by the calculated utility for the agent. Desirability reflects the value of the causal interpretation's propositions, i.e., if they would enhance or inhibit utility. Causal attribution informs an agent's responsibility for executing an action.

Likelihood and expectedness represent the probability of an state to occur and the agent's possibility to predict it. Controllability and changeability inform about the possibility to change an outcome by the agent or other agents, respectively (MG14).

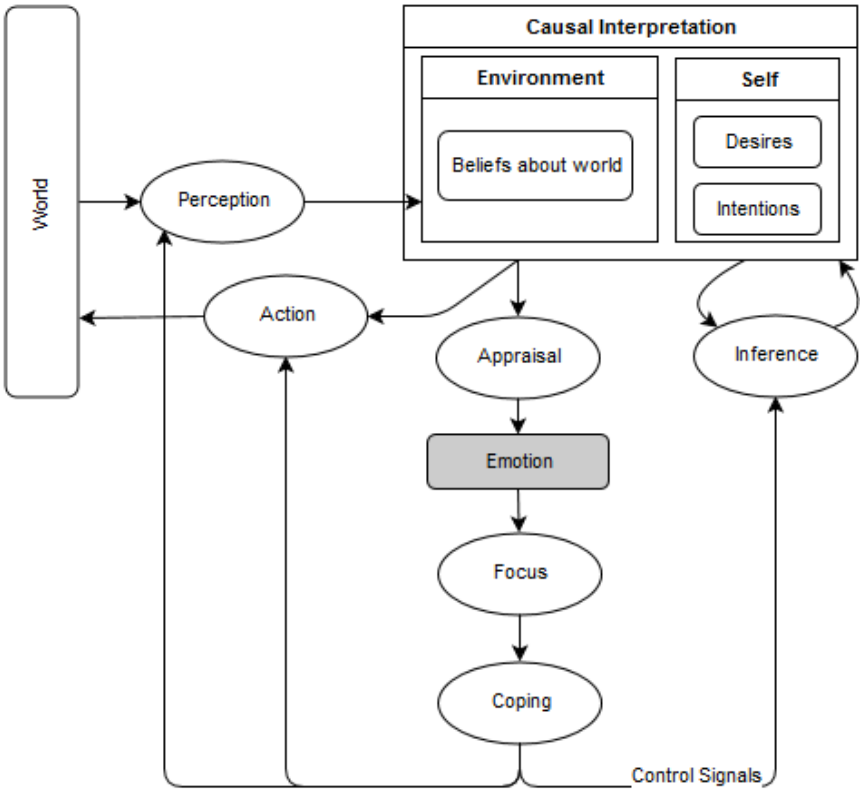

Figure 3.17: EMA's appraisal, coping, and reappraisal process (MG14, p. 62).

Each proposition in the causal interpretation is appraised with all these appraisal variables, forming an appraisal frame, which in turn can be labeled with one or more emotion terms (joy etc.) and an intensity (MG14). (This is functionally irrelevant and only used for convenience, e.g., for a mapping to facial expressions.) All appraisal frames are aggregated to the agent's mood state (a set of emotion labels). An appraisal frame's intensity and recency, together with the mood state, determine the selection of a coping strategy. Their objective is to alter the situation that triggered the corresponding appraisal by operating on its causal features (e.g., beliefs, desires, intentions, attention, and expectations). Strategies can be categorized in attention-related coping (e.g., suppress/seek more information in case of low/high controllability amongst others), belief-related coping (e.g., wishful thinking: increase/lower the probability of a (un)desireable outcome), desire-related coping (e.g., mental disengagement: lower the utility of an desired by threatened state), and intention-related coping (e.g., planning/action selection, procrastination, resignation, avoidance) (MG14).

3.4.3 FATiMA

The FATiMA (Fearnot AffecTIve Mind Architecture) model aims to provide a believable and empathic synthetic character to enhance user engagement in systems with pedagogical and training purposes (DP05). To reach that objective an agent's emotional state and personality is regarded as

key. Different to similar approaches, the agent architecture is aimed as being domain-independent, i.e., easily applicable for different personalities in different application domains. A paradigmatic case for evaluation an application of the model is given with the FearNot! project (ALD+05), providing an anti-bullying demonstrator for children. Experiments in Portuguese, English and German schools demonstrated that the children felt empathy for the synthetic characters (DP05).

In the FATiMA model, inspired by the OCC cognitive theory of emotions (OCC90), an emotion is described by its type, valence, cause (event or action), target (object or agent), intensity, and time-stamp

An agent's personality, then, is also inspired by the OCC model and given by "a set of goals; a set of emotional reaction rules; the character's action tendencies; emotional thresholds and decay rates for each of the 22 emotion types dened in OCC." (DP05, p. 131). Two goal types are distinguished: actively pursuit goals and passive goals. Emotional reaction rules represent the agent's attitudes and determine the appraisal of generic events. Action tendencies determine how an agent would react to events without reasoning about them. An emotion that is created in the sake of an appraisal process is only considered for an agent's emotional state if its potential surpasses the personality-dependent threshold for the according emotion type (e.g. anger). The nal emotion intensity, then, is calculated by the difference between the emotion's initial potential and the threshold (DP05).

To reach the objective of a domain-independent model (FATiMA modular), a generic core architecture is developed (DMP10), with functionalities and processes divided into modular independent components. One advantage is the usage of such framework for the implementation of different appraisal theories and hence contribute to standardizing appraisal-based emotion models. FATiMA modular provides a template of generic functions, with the possibility to implement them by adding specific components. The general core structure shown in Figure 3.18 enables triggering of an appraisal process by perceptions and updating the agent's memories or internal state. The appraisal process results in an affective state, which influences the action selection process. Appraisal in the core architecture can be separated into appraisal derivation and affect derivation. The former determines the relevance of an event for the agent, resulting in appraisal variables (e.g., desirability). The latter uses these appraisal variables to generate the agent's emotional state according to an specific appraisal theory. Thereby a special effort is taken to provide the basis for implementing Scherer's appraisal process of multi-level sequential checking (see Chapter 3.1), opposed to e.g. the EMA model (see Chapter 3.4.2). To reach that goal specific design principles, with the aim to generalize Scherer's requirements for its appraisal theory, are set (DMP10):

- Incremental appraisal: The appraisal process is conducted continuously, i.e., each component decides if and when an appraisal variable is generated and processed.
- Appraisal variables and related variables are accessible by all components and their dependency is specified.
- The appraisal variables and resulting emotional state are compliant with allowing reappraisal.

The introduction of an appraisal frame that integrates the decentralized access of different components to the appraisal process, together with the sketched core architecture, supports the consideration of these requirements.

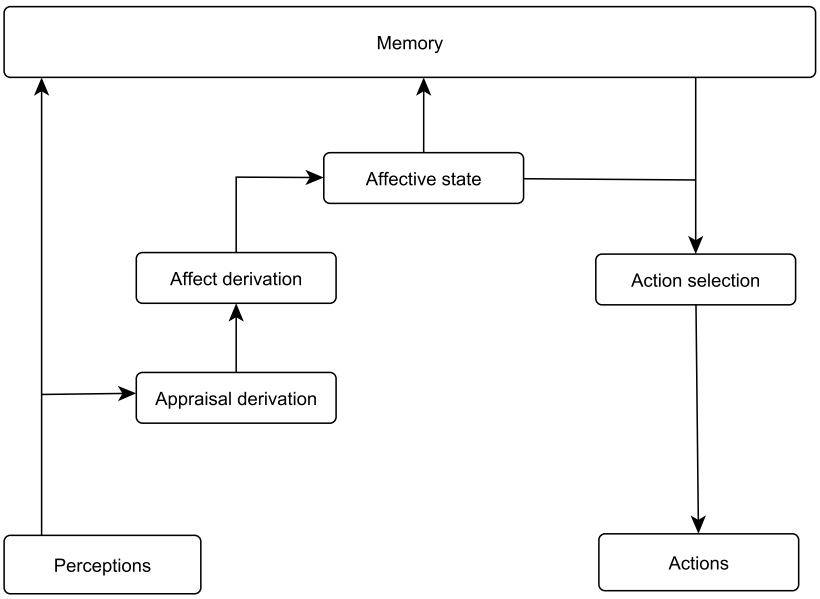

Figure 3.18: FATiMA Core Architecture (DMP10).

However, the core architecture is not sufficient and must be extended by components. Examples are (DMP10):

- The reactive component is used to generate appraisal variables by matching occurring events against a set of emotional rules.
- The deliberative component is concerned with goal-based behavior and planning.
- The motivational component models basic human needs as an additional selection criteria of goals in the deliberative component.
- The theory of mind component simulates the appraisal process of other agents.

3.5 Conclusion and Synthesis

We saw that different disciplines and different topics have to be considered when aiming to develop a model of decision making in social context. In summarizing and concluding the literature analysis, a synthesis of the different aspects from these disciplines for a model that suffices this thesis' criteria (see Chapter 2.2 and summary at the beginning of Chapter 3) will become clearer.

The different human behavior explanations in sociology and psychology can be bridged in heterogeneous agent-based social simulations (ABSS), which aim to bridge the micro-macro gap in complex systems. In doing so, both social and individual structures have to be considered "through the agent's mind'" (Cas00). A computational methodology enables to model the natural entities of social systems, including a detailed model of the mind. Only with such an approach we are able to go beyond the limits of unrealistic behavioristic rational-choice approaches, which consider humans as homines oeconomici. This approach would also enable better empirical grounding

and fine-grained policy formulation that address the mechanisms of human decision making. A possible trade-off between abstraction and specification in ABSS is to use basic building blocks of cognitive models as foundations. The detail of specification and extension, then, depends on the concrete research question at hand and on insights about emergent behavior from prototypical simplified ABSS. An example of harnessing both simplified agent-based models and mental architectures is given with a possible co-modeling with the Consumat and SiMA-C approaches. Consumat can be used as a starting point to identify cases in simulations that require further exploration (e.g., because it identifies a wide variety of possible outcomes, or unexpected social barriers for change). In such cases SiMA-C can provide more insights of agents' decision-making processes in different personalities and further specify Consumat's variables where (assumedly) necessary.

In principle, conventional cognitive architectures (as domain-independent specifications of the mind's essential structures and processes for the purpose of analyzing cognition) are appropriate to replace the simplified models used in ABBS. However, their (computational) means impacts their theory. Consider for instance production-based systems, such as ACT-R or Soar, which claim that all aspects of decision making can be represented in a production-based form (i.e., condition-action pairs), as they are the central bottle-neck and the sole mechanism for decision making. Such cognitive architectures do not base their reasoning-based mechanisms on motivations and emotions. Rather they extend their models with emotion, e.g. representing empathy with production-rules. Another lack of cognitive architectures is the dominance of explicit (function and data) representations and the neglect of an embodied generative approach. The lack of considering established theories from other disciplines (also reflected in the other reviewed cognitive architectures) is also expressed in neglecting the key principle of considering two decision making systems (considering the role of the unconscious) that interact but operate under different principles. In this regard the transition from homo oeconomicus to homo psychologicus has still not arrived in cognitive architectures and require intensive interdisciplinary collaboration.

Instead, a mental architecture is able to consider fundamental principles and their integration with high-level cognition into an holistic-unified model of emotion and cognition. Such an approach is better able to consider heuristic thinking, which focus on processing information that is adaptive to the agent's current internal and external situation. Recent approaches in behavioral economics emphasize the dominance of the automatic system in everyday decision making, in particular under uncertainty, but fail to go beyond descriptive models. Computational models can provide a functional-generative approach in modeling the underlying processes.

However, computational models of emotion primarily consider emotion as the consequence of cognition and hence prefer appraisal models without considering the embodied aspect. That is, instead of following a hierarchical evolutionary approach in integrating cognition and emotion such approaches model emotion with cognitive processes. This relates to the century-old question whether perception suffices to generate emotion or only cognition does. Of course, this depends on a clear definition of cognition. However, when separating primary, secondary emotion (or basic and extended emotion) and feelings this question is clearer to answer. Basic emotion represents bodily needs and additionally can be triggered by perception, extended emotion is a result of low-level cognition, both are processed in the primary process. Feelings are the result of a conscious perception of the bodily state resulting from emotion and are used in the secondary process for high-level cognition[23]. Opposed to that, appraisal models only consider external events as triggers of emotion and neglect low-level aspects of appraisal, which should be regarded as a layered and

[23]Hence, what William James(Jam84) regards as emotion, should be considered as feelings.

distributed process. But a computational model of emotion has also to consider the relation of emotion to other life-regulation mechanisms - as sketched by Damasio's nested model. When considering the general purpose of emotion as an adaptive and integrative mechanism bridging body and mind, internal and external world, individual and social demands, a holistic-unified model is required. The unifying information-processing function that is able to generate this functionality and serve this purpose is emotion as an indicator of an agent's state (i.e. emotion as data) that serves the function of valuation and evaluation.

4 Theories and Conceptual Model

All human actions have one or more of these causes: chance, nature, compulsion, habit, reason, passion, and desire.

Aristotle, 384-322 BC

The assertion that those psychical states which we class as feelings, are involved with, and inseparable from, those which we class as purely intellectual processes - that they form but another aspect of the mental phenomena already described; is an assertion that will appear untenable. ... Memory, Reason, and Feeling, are different sides of the same psychical phenomena.

Herbert Spencer, 1820-1903, Principals of Psychology

The first step in answering this thesis' question[1] is to find the required psychological theories in the context of our research program (i.e., the SiMA approach) and specify them into an operational form for the SiMA architecture. The usage of two concrete instances of this question - social interaction and environmentally friendly behavior - supports specifying the *representation* of these theories into a consistent conceptual model and show their feasibility of generating the expected behavior. After an analysis of the two corresponding exemplary cases, the assumed theories that fulfill the analyzed requirements and generate the case's behavior from the psychological point of view (i.e., the theories with psychoanalysts and psychologists assume underlying the behavior) are sketched in the context of the research program's framework theories (i.e., basic assumptions of the mind's functioning and how it can be represented in a mental architecture). Given the criteria of an interdisciplinary, holistic-unified, and functional-generative approach, the aim is to find the foundational elements of human decision making. These theories are then specifies and fit into the SiMA architecture. The common functionality of these theories is that they constitute means of valuations for the satisfaction of demands from different sources, which can be regarded as the core of decision making. Although an exemplification is a means to support this step, the aimed result is a generic model of basic decision making and their instantiation in a social context.

[1] ... about a foundational mental architecture that consider social context.

4.1 Exemplary Cases

To analyze the respective psychological questions and demonstrate concepts, exemplary cases are used (see Chapter 2.3). Hence, a narrative description is used to exemplify research questions and appropriate theories for them. This method support analysis of the required factors that need to be considered in answering the research question and support theory formulation. To reduce complexity, a simple story is used that still is able to demonstrate complex concepts. Especially when the foundations of the human mind are at stake, simple every-day behavior is more suitable to analyze the basic functions of the human mind. Overall, an exemplary case describes behavior of agents in a paradigmatic situation and assumptions about the generation of the behavior. This serves guiding model development and validation. However, the exemplary case serves only as a point of departure for specifying concrete scenarios in simulation cases (see Chapter 6 Simulation).

After a summary of the two exemplary cases involved in this thesis and an analysis posing sub-questions, the corresponding theories are introduced. The two exemplary cases (and their specification and simulation) were developed and implemented in sequence. Hence, the second exemplary case (about environmentally friendly behavior) was able to include the first exemplary case's gained experience and results.

4.1.1 Motivations from Nature and Culture: Eat, Share, Fight, or Flight

The original exemplary case 'Adam seeks Schnitzel' (BGSW13), which aims to demonstrate principal mechanisms for developing the basic functions of an artificial mind, is rewritten for clarification under the leadership of this thesis author by a group of psychoanalysts and computer scientists at the Institute of Computer Technology, in the context of the SiMA project into the final form (DBD+15, pp. 48 ff.). This exemplary case addresses the basic mechanisms of the human mind, with a focus on bodily needs and social norms. To further specify the model and extend it by including the topic of emotion, their bodily expressions and interpretation by other agents, a follow-up exemplary case 'Adam and Bodo' is formulated by the author of this thesis (and reviewed by psychoanalysts).

Since the exemplary cases are coherent and both are the topic of this thesis, for the sake of simplicity both exemplary cases are integrated and summarized as follows.

Exemplary case name: Two agents and a food source

Research question: How are autonomous agents' actions motivated and decided? Exemplified with the question: How does an agent cope with its bodily needs and norms in a social context?

Demonstrated concepts: Metapsychological basics of the mind: Id (drives), Super-Ego (norms), and Ego (defense mechanisms and action selection); memory-based decision making; interplay of natural and cultural factors and their integration in emotion as holistic valuation; empathy.

Context: The story consists of two agents, Adam and Bodo, and a Wiener Schnitzel[2] as a food source[3].

[2] A Wiener Schnitzel, a breaded fried veal cutlet, is a traditional Austrian dish.

[3] The reasons why the agents are in this exact situation are not considered to avoid unnecessary requirements as much as possible.

Story from Adam's perspective: Adam is hungry. *[His empty stomach signals homeostatic organic tensions that are represented as drives in Adam's mind.]*[4]

His hunger rises, which in turn led to aggressions. [Adam is an impulsive character: He cannot deal with rising drives, since he has not learned to defer his drive wishes sufficiently in his upbringing. Hence, his personality tends to fulfill drive wishes with aggressive actions. This leads to highly activating bite-fantasies with the aim of reducing hunger. However, Adam also has a strict Super-Ego, i.e., internalized norms, which forbids to be aggressive by generating an inner conflict. The Ego mediates between the two demands by converting aggression into anxiety. A share of Adam's drives is used as neutralized intensity for secondary processing[5], e.g., to reason about how the long-term fulfillment of his drives, which activates plans for searching.]

Adam searches for food, perceives a Wiener Schnitzel and decides to eat it. [Luckily this is one of Adam's favorite drive objects, i.e., the hunger drive activated memories possibilities for satisfaction. Since Wiener Schnitzel is memorized as the best object or satisfying his drives (besides the stomach drive, it satisfies the libidinous oral sexual drive) - according to the pleasure principle[6] - it is chosen as the drive object and decides to eat it.]

He approaches the Schnitzel and perceives another agent, Bodo. [Bodo activates unconscious memories of Adam's older brother, Carl, which Adam admires but also dislikes, because of brotherly rivalry in their upbringing. One of these memories consists of competing for food. Adam's perception of Bodo causes additional aggression and plans to beat him.]

Bodo is sweating and blushing heavily and shakes a little. [His bodily state, including his drive state, and the memories activated by perceiving Adam, corresponds to anxiety, which is expressed strongly via his body due to Bodo's personality. Adam interprets Bodo's bodily expression as anxiety. Based on his valuation of and relation to Bodo, this attributed emotional state increases Adam's own anxiety and anger.]

[The Super-Ego generates a conflict due to Adam's anger and the goal to beat another person. The Ego mediates by turning Adam's anger into guilt, which in turn activates plans to share the Schnitzel and plans to give Bodo the Schnitzel. Dependent on the available neutralized intensity Adam reflects on the feelings associated with the plans regarding the expected (un)pleasure gain relative to his current feelings.]

Possible outcomes: Depending on the mentioned factors, such as hunger, super-ego strength, neutralized intensity, memories, bodily expressions, etc., Adam chooses the action that subjectively provides the highest pleasure gain and the lowest unpleasure gain in the long run. Possibilities are to eat the Schnitzel alone, to share it, to give it to the other agent, or to avoid the situation and to leave.

Related sub-questions: see Chapter 4.1.3.

Assumptions: Formulated in chapters 4.2 and 4.3.

[4] Italic text in brackets describe the explanatory psychic processes underlying the preceding behavior description.
[5] for a description see Chapter 4.2.1 and 4.3.2.4
[6] for a description see Chapter 4.2.1

4.1.2 Think about Environmentally Friendly Actions! It is a Good Feeling to Conform and Follow Your Norms

To analyze the requirements of a decision making model applied in the area of environmentally friendly behavior, the exemplary case of getting a message from a friend about switching to green electricity is chosen.

Exemplary case name: Social Prompting

Research question: What motivates environmentally friendly behavior? Exemplified with the question: How can a message in a social network address motivations to switch to green electricity?

Demonstrated concepts (additional to 4.1.1): Social emotion, social norms, imitation and peer pressure.

Context: Caroline and Victoria are social media friends. They are similar regarding their age, profession, and interests. Both share the same peer group.

Story: Victoria is currently at work. She is a little anxious, tired, and hungry. Therefore she browses on a social media platform, where she recognizes that her friend Caroline switched to green electricity: 'What a good feeling: I finally switched to green electricity. To a green future for us all. Yours, Caroline.' [Victoria knows about the advantages of green electricity and follows the norm to protect the environment if necessary. However, switching to green electricity was never a serious topic for her. Now, that she got interested in the topic she thinks about switching to green electricity, since the message from her friend, directly and indirectly, emphasizes how good green electricity is, which is compliant with Victoria's personality and experience. The message activates Victoria's memories about green electricity and her norm to protect her environment through different paths.]

Many shared friends of Victoria and Caroline comment on the message and show their approval. Victoria unconsciously is envious that Caroline gets the attention from so many friends and that Caroline fulfills shared values. Her unfulfilled bodily needs (due to prohibition to eat and rest at work) are sources of conflict and are sublimed - due to their similarity - to get similar attention as Caroline (since it is feasible to do - opposed to eating - and Victoria and Caroline are similar). This conflicts with her norm not to be envious, especially with friends. The conflict is mediated by identifying with Caroline, which leads to adopting Victoria's goal to switch to green electricity (to protect environment) by imitating Caroline's behavior.]

Possible outcomes: The message contains web links to get further information and to order green electricity products. Depending on the determinants (see Chapter 6 Simulation) and their impact on the decision factors, Victoria decides to click the link or continuing working.

Related sub-questions: see Chapter 4.1.3.

Assumptions: Formulated in chapters 4.2 and 4.3.

4.1.3 Exemplary Case Analysis

The exemplary cases support analyzing the conceptual requirements of the research questions at hand. Next, by explicating the implicit requirements sub-questions are formulated and an starting point for appropriate concepts that should be considered in modeling.

Which factors motivate human behavior and which conceptual mechanisms can explain the causal chain from trigger sources to action? This question implicitly is concerned with the very basics of the functioning of the human mind. A foundational answer has to start from basic functionalities and demonstrate how other functionalities use them, hence considering evolutionary development.

Analysis of the Exemplary Case 'Two agents and a food source'

The most basic question (that drive the evolutionary development of living beings) is that for survival and reproduction - bodily needs - sourced in the structure of our body. Different to other animals, control structures for satisfying these needs are not solely bodily (e.g. osmosis). Evolution has largely outsourced such implicit control from the targeted body parts - by creating an explicit control system - to the brain[7]. Still, most of the control happens automatic, on this level termed unconsciousness. The homeostatic principle prevails on this level, but operates with other means (informational instead of physical), which are better describable on a more abstract level - the psyche.

How can we represent the generation and processing of such bodily needs from its physical source to its action via the human control system, i.e., considering the separation of source and control? This is the central question of autonomy. The concept of drives (see Chapter 4.3.2) provides a basis for an answer.

Adam is hungry (i.e., has a bodily deviation) and everything that will happen from now on, is triggered and caused (directly or indirectly) by that bodily deviation. (Even) humans cannot escape nature. But humans are the first animals who evolve culturally, not only naturally[8]. Hence, natural demands are extended and adapted to cultural demands. With nurturing and intentional teaching, internalized norms become part of human motivation. However, humans are still physical beings and have to relate all motivations to their body. Or in other words: the substrate of culture has to be (directly or indirectly) nature.

How to represent this kind of embodiment? And how to join the different sources of motivations (drives and norms) in decision making? This question implies the need of a framework to integrate different factors and aspects in human decision making. A point of departure for this answer are models provided by psychoanalytic metapsychology (see Chapter 4.2.1). In this framework all processing in the psyche is driven by the demands of drives (including their deviation by internalized norms).

Adam perceives a Schnitzel to satisfy his hunger. How does he know this object? One central assumption is the requirement for memories. But why does he have memories? Primarily to know how to use the external environment for the satisfaction of demands. First and foremost Adam does not care what this perceived thing is called. It activates memories about similar things and information about their purpose for him. Hence, memories primarily inform how to satisfy needs (and based on them, goals). Thus, perception serves satisfaction of needs. But not only external perception activates memories, but also drives activate memories as experienced satisfaction.

Hence, memories serve as a basis of how to deal with our internal world in the perceived (and hence interpreted regarding the internal world) external world. How should memories be formed to serve this purpose and which information should they provide? The minimal information

[7] We can observe that technology chose a similar development: From mechanical-based control to information-based control.

[8] The term culture is used in its original etymological broad sense, describing everything (material and non material) humans create artificially based on natural resources. Regarding the human mind and its manifestation, nature is sources in genes, culture is sourced in nurture.

required from memories is how to relate internal needs (and goals) to external affordances, i.e., to value the external world regarding the good and bad effects on the agent's goals.

Adam's hunger activates memorized possibilities for satisfaction. But which memories are activated, how could they be represented; as abstract, semantic information; as plans, as context? If memories of a specific Schnitzel are activated, which information do they provide? The minimal relation of the external to the internal world via memory is the relation of their minimal representations in the mind: objects and drives, e.g., how good an object satisfied hunger in the past. But objects are not only able to satisfy drives, they also may increase drives or bring direct harm via pain. The drive concept is not enough to account for the latter, and has to be extended by a concept of emotion. Hence, perception and memories are not only to approach exterior, but also to avoid it.

Hence, the essence of memories is to form expectations about the effects[9] of things, providing information about their value for the agent. In this regard memories are anchored with affective principles. But the value of a thing is measured based on the current demand. Hence, memories have to inform the ability of things to reduce current demands and their risk to bring harm.

But Adam does not only perceive an isolated Schnitzel, he also perceives Bodo, which is memorized as similar to Adam's brother Carl. How we treat people is dependent on what we associate with them. And since Adam knows much more about Carl than about Bodo, he unconsciously treats him mainly in this regard. We are only able to act according to our experience and knowledge. This again emphasizes that we construct our world subjectively. The requirement for associativity, spreading of activation and valuation in our model to account for such functionality become stronger.

As before Adam uses his memories to relate inner and outer world. Again, Adam's memories may inform how good Bodo can be used directly to satisfy his drives. But Bodo is not only another object, but also a subject (since he expresses cues - blushing etc. - of an internal subjective state). Are Carl and Bodo memorized as supporting satisfying specific drives or as hindering their satisfaction and how can this be represented in memory? How is the relation, e.g., the dominance and admiration mentioned in the exemplary case, to Bodo represented in memories? What indicates if Bodo is an agent to cooperate or not? In any way, memorized valuation about Bodo has to subsume all experience with him.

Adam perceives that Bodo has bodily expressions, such as sweating and blushing. These are indications that he is an agent just like him and he unconsciously uses the concept of emotion to attribute an internal state to Bodo, which in turn affects Adam's own state, considering it as an additional aspect of valuating Bodo.

Another aspects to consider in valuation, especially when the object is also an agent, are valuations regarding long-term effects. This is especially relevant for valuing the next level of perceiving the external world - situations. Its valuation, i.e., relating the whole situation to the agent's inner state, is often different than the sum of valuation of the situation's objects. Context matters in valuation. Another aspect is the consideration of different, possible contradicting, valuations regarding different needs. For instance, an object may satisfy one drive, but dissatisfy another one, or cause pain. Pleasure and unpleasure must be considered commonly. Emotion is a concept that accounts for these requirements.

[9] Compare the etymology of 'affect' with 'effect', both having 'facere' - to do - as basis. Hence the etymology can be interpreted as dependence of the mind's activity on affect.

Adam's drives and memorized associations of the perceived situation lead to anger. Adam's super-ego pose rules that may be conflicting, e.g., the prohibition to be angry. Additionally, the outer world does not only have to be related to demands from drives but also to demands from norms, represented by the concept of a super-ego. Hence, a concept is needed that merges the demands form drives and super-ego, which may be conflicting, in their valuation.

Overall, memories have to relate the direct and indirect effects of the outside world to the agent's inside world, represented by drives and super-ego demands (to consider both, nature and nurture), under consideration of their context. Hence, to consider the whole picture, an integrative holistic concept to represent the internal state is required. A concept of emotion is able to consider these aspects. But how are the concepts of drives and emotions exactly related and how do they work together in their valuation? For instance, is emotion a subsumption of drives and their valuation and how do drives and memories lead to Adam's anger?

Adam's Ego mediates between super-ego and drive demands, which results in having guilt, which in turn activate new memories and plans. This is an example of how the concept of emotion integrates super-ego demands, drive demands, and perception. This enables choosing a plan that considers all theses aspects. However, parallel to this holistic consideration, activation from the single factors are still considered.

If Adam would act only by following the activated memories, he would live completely in the past. But this is only valid for the domain of the unconscious - the primary process. However, unconscious data is partly used in the secondary process to reflect on the primary processes' decisions, if necessary and possible. Hence, activated memories are only a basis that has to be reflected on, in particular, they have to be adapted to all current decision factors. This includes the agent's ability to separate his current from the expected state, which is not possible in the primary process (since it does not consider causality and time). Even if the external situation would be exactly like the memorized one, the agent's current state is unlikely to be exactly as the past one. Secondary processing also includes the recognition of other agents' state and reasoning about their intention.

This demonstration of a requirements analysis and concept description addresses the most important questions and concepts. However, in the process of further analysis and development, these questions and concepts are specified and extended in the remainder of the thesis at hand.

Analysis of the Exemplary Case 'Social Prompting'

The exemplary case that examines concepts for environmentally friendly behavior addresses the same basic questions as the previous case. As emphasized, the two cases demonstrate different instances of the same question. However, opposed to the former exemplary case, this one demonstrates how the same concepts are used as a basis to generate more sophisticated decisions, at least from a superficial view. Additionally, high-level decisions, such as the one at hand about environmentally friendly behavior, are also driven and impacted by low-level factors. In particular, decisions are also impacted by factors that seem to be detached from them. For instance, at first glance it may seem counter-intuitive that the current bodily state has an impact on the decision to switch to green electricity.

Even if the concepts are the same to tackle the questions in both exemplary cases, the model and implementation requirements differ. The reasons on model level can be summarized by (1) the different requirements of an model for an Artificial-Life simulation (simulating behavior, as in case of the former exemplary case) and of a decision-making simulation (only simulating decisions, as in case of this exemplary case) (2) the different requirements of a test-bed for translating concepts

about the human mind into a computational representation and of applying an analyzed model for subsequent questions. That is, starting with basic research, the usage of a conceptual model in Artificial-Life simulations is necessary to translate the concepts into a computational model to test and explore them. Only after that, we can identify the functional essence of these concepts. This also enables an integration of the different concepts into an unified and parsimonious form (especially if the model is 'naturally grown', as the SiMA model).

Hence, for the first exemplary case a conceptual model is developed and integrated into the SiMA model, and simulations are conducted using an implementation of it. For the second exemplary case the SiMA-C model (see Chapter 5) as a derivation of the conceptual model is developed, under the consideration of the mentioned requirements and integration possibilities, which are enabled by an analysis of the previously developed conceptual model and its simulation.

However, additional to a different usage of the same concepts, the question at hand about environmentally friendly behavior requires some specializations and extensions of the basic concepts used in the former exemplary case. In the remainder of this chapter the analysis of the former exemplary case is only extended regarding the new aspects. The decision to act environmentally friendly has many causes, and usually only the interplay of these causes suffice for a clear decision. Even if - dependent on the personality - different causes may prevail, seldom is a cause so clear and pervasive to allow for a mono-causal explanation. As mentioned, even bodily causes may come into play: The ability to reflect on the current situation is indirectly dependent on the agent's drives (see Chapter 4.3.2.4), and the valuation from drives that are forbidden by super-ego rules (i.e., norms) may be used for other goals. Besides past experiences with the topic of green electricity, different internalized norms, compassion and identification with nature, but also social pressure and imitation influence the decision. In the end its is usually an interplay of these factors and often a conscious reflection of them - even if the data that is reflected on is caused unconsciously. This also holds for the exemplary case, where the message activates Victoria's memories about green electricity and her norm to protect her environment through different paths.

The exemplary case generally emphasizes the role of activations. Attitudes, e.g., in the form of valuated memories (experiences and norms), do not have any impact if they 'do not come to mind', i.e. are not activated. In the exemplary case, Victoria knows about the advantages of green electricity and usually follows the norm to protect the environment. However, this attitude only has an impact on her decision by making getting green electricity as a focused topic to think about and connecting it to her norms and experience. Moreover, the same norms and experiences with a topic can lead to different decisions by different prompting messages. In the end, it is always about the interplay between the inside and the outside situation. Following our principles, Victoria's attitude is depending on how they get activated (by the external environment). Not only the activations by the current situation are relevant, but also the previous ones. Victoria is at work, so the goals she previously and currently follows and their relation to the memories activated by the new situation come into play as well.

But it is not only about the data that is activated by different means. The social context is crucial. Decisions seldomly are done outside social context. And even if physically alone, the social influence is mentally internalized by norms and other memories. Victoria gets the message from a good friend with similar personality and hence the message is focused due to the identification with Caroline. The appreciation of Caroline's message by their peer group increases the demand for conformity and social pressure.

How relevant norms and experiences are activated is not only influenced by the social context, but also by the emotion transported by the prompting message. Caroline emphasizes that it is a good feeling to switch to green electricity. This not only enables Victoria to emphasize with her, but also impacts her own emotional state and hence her decision, since the overall objective of a decision is to subjectively enhance - in a short or long term - the person's state.

Even if humans are programed to emphasize with other humans, we are also able to emphasize with nature, given a strong relation and identification with it. Victoria knows that she is a part of nature. Hence, perceiving nature in a state of unpleasure also affects her.

Overall, all these reasons have to be integrated into a unified model of decision making to demonstrate how their interplay determines an agent's decision.

4.2 Framework Theories and Basic Assumptions

Before specifying the required concepts that are identified in analyzing the exemplary cases, a condensed description of the used framework theories - supporting holistic conceptual modeling - is given. This theoretical framework is a good point of departure and serves as a glue in specifying concepts that are suitable for a computational model of the mind. These theories are extracted from the SiMA principles and a literature analysis (see Chapter 3).

4.2.1 Metapsychology

The core of the used framework theory is given by Freud's meta-psychological models of the mental apparatus, in particular as interpreted in the SiMA project. One of them, the so-called structural model, represents the *functional structure of the mental apparatus* with three functional units: Id, Super-Ego, and Ego (BDZ+13). The Id represents bodily demands as drives (demands from the body for the mental apparatus), the Super-Ego represents internalized demands from social rules, and checks the compliance of mental processing with them, the Ego mediates between the different demands by considering the affordances of the external environment (BDZ+13).

The three functional units operate in different *processing modes*: primary and secondary process. The processing mode defines the degree of structure and processing methods of data. Data in the secondary process have a high degree of symbolization and is associated in a hierarchical and causal way (sequence, time etc.), which is not the case in the primary process, where data structures are still flexible and the valuation of data can be displaced and condensed using its associative form, but no negation is processable. Data processed in the primary process are unconscious, and in the secondary process they are pre-conscious or conscious (BDZ+13). The data structure of the primary process is called thing presentation, which is extended with according word presentations in the secondary process. The Id and the Super-Ego operates solely in the mode of the primary process, the Ego process in both modes.

Two *processing principles* can be distinguished. The pleasure principle states that all psychic processing in the primary process aims at immediate and maximal satisfaction of demands (from Id and Super-Ego), the reality principle states - complementary to the pleasure principle - that the secondary process considers the reality given by the external world in reasoning about the best long-term plan to satisfy these demands.

The three functional units use *methods that follow these principles*. The central method of the Id is cathexis: Homeostatic differences in the body generate drive tensions that are represented in the mental apparatus as quota of affect, which cathect memories that are associated with satisfying these tensions (SHDD14). This cathexis may be changed in the primary process by the methods of condensation and displacement. The amount of current quota of affect is termed unpleasure. After executing a cathected action, quota of affect is reduced and corresponding pleasure is generated, which - in combination with the currently existing unpleasure - can be described as basic emotion (SHDD14, p. 21). Demands from Id and Super-Ego can be conflicting. Additionally, conflicts may occur between demands and reality, but also within a functional instance (e.g., conflicting drives in Id, conflicting norms in Super-Ego, conflicting norm with ego-ideal in Super-Ego, conflicting plans in Ego) (Lis09, pp. 108 ff.). The Ego uses defense mechanisms (Lis09, pp. 91 ff.) as a method to mediate these conflicts. These defense mechanisms may use condensation and displacement. Another key method of the Ego to comply with the reality principle is drive delay: Instead of handling demands in a reactive way, their short-term fulfillment is delayed - due to an impeding situation - to increase their long-term fulfillment (BDZ+13). The available unpleasure can be regarded as a measure of the intensity of processing of the Id, a personality-dependent share of this unpleasure as so-called 'neutral intensity' is a measure of the ability of processing data in the Super-Ego and Ego. All these methods can be described by three points of view or perspectives (Lis09, pp. 74 ff.). Cathexis and its adaption is regarded as the economic aspect of processing, conflict generation and mediation represent the dynamic aspect, and the topic aspect is given by the different processing modalities in the mental apparatus' three functional units.

4.2.2 A State Indicator for Life-Regulating Mechanisms

The key control principle in living beings can be regarded as homeostasis, for both the physical and psychic level. Different homeostatic mechanisms of life regulation evolved incrementally: reflexes, metabolisms, drives, emotion, feeling, reasoning (Dam03, p. 37ff.) (for details see Chapter 3.3.2.3). With advancing evolution these mechanisms are refined for the sake of better adaption on the environment. This extension of homeostatic mechanisms can be regarded in a nested form (Dam03, p. 40ff.). One metaphor that demonstrate the integration and relationship between the different mechanisms is an irregular grown tree, with new evolved mechanism as new branches of old trunks.

In this scheme, emotion is regarded as the top-most homeostatic mechanism for life-regulation that is still grounded in the body. The key role of emotions in decision making is given by its evaluative function (NSB06). The embodied aspect of this process is emphasized by translating the evaluative result of the psyche into an a bodily state and (by that) bringing the current state, and hence the decision to consciousness.

Overall, emotion can be regarded as grounded and embodied cognition. In particular, emotion are expressed bodily and mentally, and hence demonstrates the unity of body and mind. This is in line with (Sol96, p. 494), stating that an affect can be triggered from the internal and the external world and has simultaneous somatic and mental manifestations, which are two different kind of representations of the same underlying unconscious process. Solms uses "affect" as the primary sensor modality[10] for perceiving the subjective inner world, which provides material for consciousness (Sol96, p. 495). The qualitative grading of this internal perception is given by a

[10] Solms compares this with our five sensory modalities for perceiving the outer world (Sol96, p. 495).

measure of pleasure and unpleasure (Sol96, p. 496). The combination of these grades with other perceptual qualities is the source for "complex emotions" (Sol96, p. 496), such as mourning.

This view is compliant with our general two possibilities of self-perception: as an object and/or as a subject (Sol96, p. 495). In this regard emotion is the mental representation of the underlying affect (Sol96, p. 495). Both ways of perception represent something that is unconscious. In this regard emotion cannot be regarded as a mental representation of a somatic state, nor as an inner perception of an external event (Sol96, p. 495).

Instead of focusing on terminological discussions and folk psychology concepts of emotion labels as often done in emotional theories (see Chapter 3.3), a computational approach has to focus on the information-processing function that emotion serves in a simulation model of the mind. However, for the social aspect of emotion, their bodily expression (as a basis for empathy) are as important as their role in decision making. For such an approach, a dimensional representation of emotion using its building blocks, and its bodily expression is beneficial. These building blocks are fed from bodily needs, norms, and perception, and hence integrate these factors and provide a holistic means to indicate the agent's state - indicate it to the outside world (to other agents) as bodily expressions and to the internal world (to the mind) as feelings. Hence, this state indicator can be considered as a meta-data about the agent's state for the sake of integration (see above), processed by different life-regulating mechanisms. Processing such form of representation and its bodily expression is able to serve the functionality as described in Chapter 3.3.2.2, i.e., motivation, decision making, higher control, appraisal, and empathy, which will be shown in the remainder of the thesis at hand.

4.3 Specified Concepts

The first analysis of the research questions at hand using exemplary cases (see Chapter 4.1.3) served as point of departure to identify sub-questions, the required concepts, and the factors that should be considered in answering the research question. After framework theories and basic assumptions for the required concepts are sketched in the previous section, the next step is to specify concepts with the goal to integrate the conceptual model into the SiMA model, on a sufficient level for computational simulation. This requires further analysis, amongst others: How exactly do these abstract concepts operate in the mind and how can they be translated into a computational model? Which aspects of these concepts should be represented as data and which as functions? As described in Chapter 2.3 this step is supported by the specification of simulation cases, which are described in Chapter 6.

4.3.1 Valuation: The Core of Decision Making

An analysis of the exemplary cases (see Chapter 4.1.3) shows that psychic processing serves selecting the goal (i.e., an action and object) that is best adapted to the current internal and external situation. The means for this objective can be described as valuation. Before describing the concrete valuating concepts, valuation is introduced as an umbrella term, derived from the framework theories in Chapter 4.2.

Different mechanisms to generate goals in a given internal and external situation evolved in humans. Given the need for a natural basis, these mechanisms must be directly or indirectly

anchored in the body. The basic source for these mechanisms is a lack that generates a need. The bodily source and expression of this lack and the feedback of its fulfillment can be termed affective. The mechanisms that evolved to act on this bodily state can be described as valuating goals that are associated with this state of lack or fulfillment. Valuation, then, is attributing value to a goal. Originally a goal's action has given direct value to the body, i.e., to its preservation, survival, and reproduction. In the process of evolution, derivations of goals that brought indirect value to the body were developed (e.g., working for satisfying hunger) through acculturation. Hence, with nature as the substrate of culture, psychological demands (e.g., norms, ideals) arose based on bodily needs. That is, even if they seem to be abstracted from the body, they originate in it. In the end, humans 'feel' through their body how good and bad a goal is expected to be. So, goals may be valuated for satisfying different demands, direct bodily demands and indirect psychological demands.

Overall, activation and valuation of relevant psychic content for meeting demands in the environment's affordances is the purpose of the human's evolved control system. All activity directly or indirectly corresponds to this purpose. Psychic intensity (i.e, the intensity of psychic processing), then, is dependent on the intensity of demands. To meet these demands, relevant memories are activated. A crucial purpose for humans to store experience, i.e., to have a memory system, is to know the value of things and actions. Depending on their memories, goals are associated with demands they directly or indirectly fulfilled. But memories are not only activated by demands, but also by external perception. The memorized association to valuations, now, gives further information for already activated memories and may activate further memories, and adds to the available psychic intensity. Hence, the current state is given by psychic intensity from demands and activated memories, and used for modulating the valuation of psychic content, i.e., perceptions and memories. Hence, the current state regulates the relevance of memories, and thereby enables adaptation on the given internal and external situation. In this regard it serves as a valuation mechanism.

Different concepts are used to represent humans' valuation mechanism, e.g., drives, emotions, feelings, and reasoning. These concepts can be regarded as incremental representations (or indication) of humans' current state and its mechanisms. These mechanisms, then, evolved by adapting on the growing aspects of the evolved demands (bodily, psychological, social). Following a general principle in evolution, these mechanisms incrementally shifted from pure bodily mechanisms (reflexes) to psychological mechanisms (reasoning)[11], since it proved beneficial for their purpose: To indicate the current state and provide (valuation) mechanisms to enhance this state, reinforcing these actions, by which an adapted agenda of survival in a given environment is enabled.

Overall, together with activation, valuation can be regarded as the foundation human information processing. In this regard information can be regarded as valuated data. Hence, valuation is an interpretation of how data can be used for the sake of the system, i.e., the person. Only valuation gives data meaning and informs about possibilities and reasons for action.

4.3.2 Drives

A central question in any system that is supposed to act in pursuing its own agenda, is about the source of its autonomy and motivation. Associated questions are: Why does an autonomous

[11] A similar pattern can be seen in engineering, where since the advent of computer technology mechanical processing is shifting incrementally to information processing for more flexible control.

agent perform any action at all? Which mechanism is the source for such actions? How does the agent regulate its internal state? The psychoanalytic concepts of drives provide a point of departure to answer these fundamental questions: The source of all motivation lies in bodily needs - self-preservation needs (such as hunger) or sexual needs (such as intrinsic sexual pleasure when eating). Later in mental processing, these drives may be adapted by super-ego rules (social norms).

(Deu11, p. 79ff.) describes the history of the drive concept's usage in the SiMA project. Generally, the concept is used to represent and measure a bodily demand, which the mental apparatus should fulfill. Hence, a drive "is a process within the body, which produces a state of tension, which directs the organism towards an aim"[12]. The mechanism for directing the organism is to valuate possible goals from memory. Hence, the questions of why an agent is motivated and how it fulfills the motivation are grounded in the same mechanism, since drives are a means for motivation and valuation.

4.3.2.1 Drive Generation and Representation

The existing concept and implementation of drives (Deu11, p. 79ff.) was refactored in the thesis at hand for further translation of the psychoanalytic knowledge into a computational model. This includes re-conceptualizing the function modules in the SiMA drive track (see Figure 4.1), refactoring the drive mesh data structure, refinements in the generation of drives, the introduction of sexual drives, the consideration of a feedback mechanism (triggered by drive fulfillment), multiple drive satisfaction by one action or object, and the extension of drives' valuation mechanism (implementing the psychoanalytic concept of hallucinatory wishfullfilment). In the remainder of this chapter this enhancements and changes of the drive concept are described.

The source of a drive is homeostatic imbalance in a bodily organ, resulting in a *continuously* flowing tension[13]. This distinguishes drives from other bodily stimuli, which occur *discretely* or stem from outside. Most importantly this excludes the concept of pain to be represented by drives. In the SiMA architecture self-preservation drives and sexual drives are distinguished, with the hormone system (represented by the psychoanalytical concept of Libido) and erogenous zones (oral, anal, genital, phallic[14]) being the bodily source for the latter. In the process of development, sexual drives, which are originally associated to self-preservation, are detached from homeostatic needs and strive for satisfaction independent from them[15]. Following the specification of intrinsic motivations as leading to activity for its inherent satisfaction (OK07), sexual drives can be regarded as an end in itself, i.e., pleasure gain from satisfying sexual drives is independent from the satisfaction of bodily needs (SDW+13). Hence, sexual drives are conceptualized in the SiMA approach not only for the purpose of reproduction, but as a general source of intrinsic motivation (e.g., for exploration and curiosity, cf. Panksepp's Seeking System (Pan98, p. 51)).

A bodily imbalance (see function modules F39, 40, 1, 2 in the neural layer and neuro-symbolic layer in Figure 4.1) is signaled to the mental apparatus as a drive tension. This physical value

[12]Source: SiMA Glossary, http://sima.ict.tuwien.ac.at/description/
[13]Source: SiMA glossary, http://sima.ict.tuwien.ac.at/description/
[14]Freuds metapsychology theory distinguishes four so-called partial sexual drives. The sexual drive's quota of affect as representation of libido - i.e., hormones - is split in the four partial drives (oral, anal, genital, phallic), which subsequently form drive representations. The ratio of splitting depends on the agent's personality, in particular how she undergoes the four development phases of childhood that metapsychology assumes.
[15]Source: SiMA glossary, http://sima.ict.tuwien.ac.at/description/

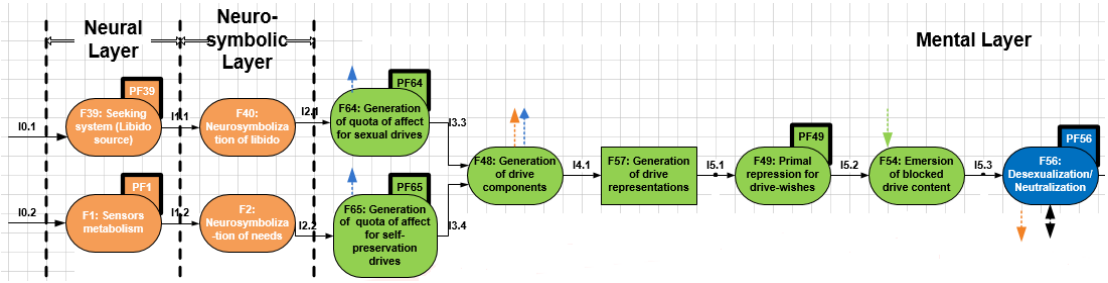

Figure 4.1: Drive track in the SiMA model (version 38s) as adapted by the author of this thesis (DBD+15, p. 83).

is represented as the information value 'psychic intensity' (SDD14), which is used as 'quota of affect' to valuate memories that are associated to fulfilling the drive (see F64 and F65 in Figure 4.1). The mental representation of the drive source (e.g., stomach) and its corresponding quota of affect form a drive candidate. Depending on the agent's personality the bodily need should partly be fulfilled by aggressive and partly by libidinous actions. Hence, the quota of affect is split into a libidinous and aggressive part. This lead to splitting of one *drive candidate* into two *drive components* (see F48 in Figure in 4.1).

To fulfill the generated drive component, memories who are associated to their satisfaction are activated and thereby valuated using the drive's amount of quota of affect [16]. The intensity of how this activation is spread in associated memories is determined by the drives' quota of affect. Overall, this process results in extending the drive candidate with a drive aim (represented by an action that decreases the organic tension) and a drive object (with which the drive can achieve its aim), and thereby forms a *drive representation* (see F57 in Figure 4.1, resulting drive representation: see Figure 4.2).

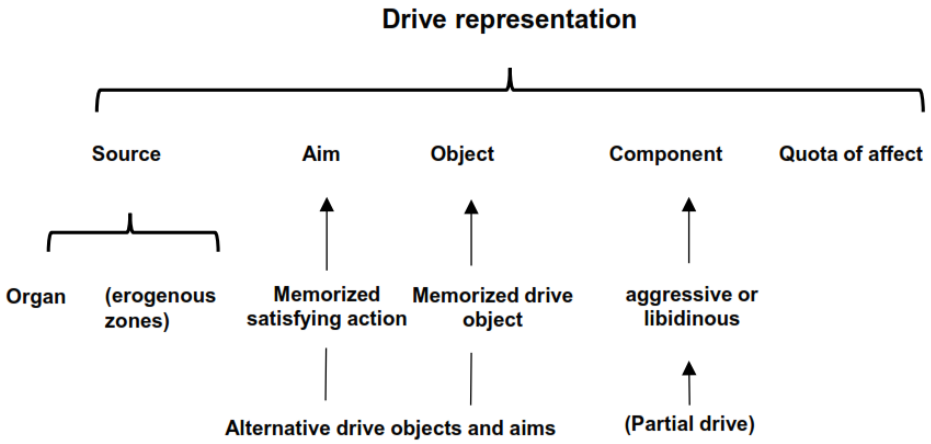

Figure 4.2: Drive representation as mental representation of a bodily need. The quota of affect indicates the homeostatic tension of the bodily source, which may be an organ in case of a self-preservation need, or an erogenous zone in case of a sexual drive. The drive component determines if the drive can be satisfied by an aggressive or libidinous action that fulfills the drive aim upon an drive object. Alternative memorized drive objects and aims are additionally associated.

[16] This process is called cathexis in psychoanalysis (see Chapter 4.2.1).

The valuation criteria of drives is given by the pleasure principle, i.e., maximization of pleasure, which is gained by recognizing the satisfaction of a drive (see Chapter 4.2.1). Drive objects and aims that were best valuated in the past (i.e. brought the highest amount of satisfaction) are valuated best. The valuation done by cathexis is based on (1) the memorized valuation and (2) the available quota of affect. Memorized valuation is represented by an association between a memorized drive representation and an memorized object and action. The valuation process consists of the following steps (SDD14): (1) Activation of memorized satisfactions of the drive with the search criteria of drive source and drive component using an associative search algorithm, (2) Assignment of a share of the drive representation's quota of affect to the data associated to the memorized drive satisfaction (i.e., assigning a value between 0 and 1 to memorized object-action pairs), (3) Consideration of accumulation by multiple valuations. The overall process (in psychoanalytic terms called 'hallucinatory wishfullfillment') results in weighted association of the drive representation with possible objects and action that are able to satisfy the drive. The best expected object and action is chosen as the drive object and aim, with the possibility of changes in further processes (e.g. due to high effort to get the wished drive satisfaction).

4.3.2.2 Integrated Drive Object Categorization

This valuation process can be described as valuating memorized and perceived objects. Since the valuation process categorizes how these concrete objects, also called exemplars since the represent specific examples of categories (Mur02, p. 49), may satisfy an agent's drives in an integrated way[17], the process is termed integrated drive-object categorization (Sch12). The integration of internal demands and external affordances is done by an activation framework. Hence , an activation-based approach is used to find appropriate memorized exemplars that may be used to satisfy current drives and are realistically reachable. For the activation of appropriate memorized exemplars by drives and external perception, an associative search algorithm is used, which enables a directed memory retrieval of appropriate exemplars. In the activation process the categorization criteria similarity (in case of external perception) and expectation (in case of wishful satisfaction of drives) are considered and aggregated. The activation functions consider how good an stored exemplar would satisfy all current drives. This expectations-based activation corresponds to top-down perception (Pyl99). An exemplar that is able to satisfy all drives best, would get the highest possible activation. The similarity-based activation by physical attributes corresponds to bottom-up perception (Pyl99).

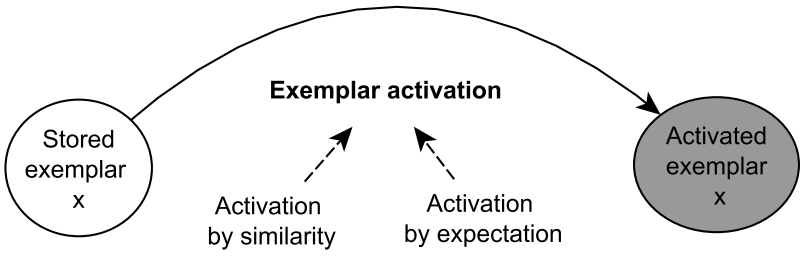

Figure 4.3: Memorized exemplars can be activated as drive objects by two paths: By similar exemplars or by expected drive fulfillment (SWB13).

[17] The process integrates activations of an exemplar from different sources, i.e., expected multiple drive satisfaction with the affordances of the external world

The whole process can be described as graded category membership of possible drive satisfaction, where multiple category memberships are possible (see Figure 4.4). That is, the degree of category membership corresponds to the amount of satisfaction that the object is expected to bring, and hence corresponds to the degree of valuation. The categorization is accomplished by an integrated multi-criteria activation process (with expectation-based and similarity-based criteria). For details see (Sch12, pp. 69 ff.).

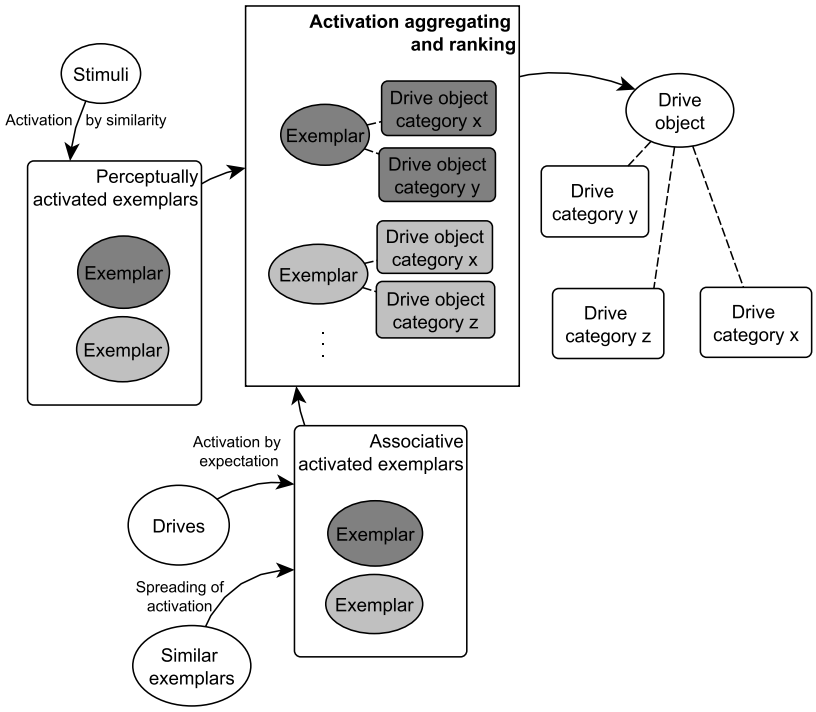

Figure 4.4: Integrated drive object categorization Using an Activation Framework (Sch12). Exemplars (memorized objects) may be activated via similarity from perceived objects or via association with drives (dependent on the association weight and the current drives' quota of affect) and similar exemplars (dependent on the association weight). Activation from different sources is aggregated and the drive categories (i.e. drives that are expected to be satisfied) are associated to the drive object. Category membership (i.e. valuation of the object by the drive) is depending on the exemplars activation (depicted by shades of grey).

Overall, this process results in activating and valuating memories that are directly or indirectly associated with satisfying the need behind the drive. From these valuated objects and actions an agenda is formed. The valuation of this goals may be changed by subsequent valuation mechanisms, such as emotion, using other valuation criteria.

4.3.2.3 Drive Cycle

By valuating objects and actions in the process of generating a drive representation, valuated goals are formed. In further processing the valuations may be changed, e.g., due to changing the valuation of socially unaccepted drive objects in the defense mechanisms ((SWJ+14)). Following the separation in strategic and tactical goals (PP10), the resulting goal can be regarded as a strategy for which a tactical plan is generated in the secondary process of the SiMA model (for

an overview of plan generation in the secondary process see (WGF+15)). However, these strategic goals are always derived from physiological drives and hence are grounded in the body. Thus, accomplishing a goal also satisfy underlying drives.

Eventually in the process of goal execution, the agent executes an action. The feedback of the drives' fulfillment by this action leads to the generation of pleasure, which indicated how good drives were satisfied. This can be regarded as feedback in a control loop (see Figure 4.5) and can occur in two ways: (1) The execution of an action may feed back on an organ, changing organ tensions, and hence subsequently reduce the generation of quota of affect. For instance, eating alters the state of the stomach, which in turn changes the drive's quota of affect. The difference in the drive's quota of affect is then calculated as pleasure (see orange arrow in F48 in Figure 4.1). (2) In case of actions that do not produce direct bodily feedback, perceiving the fulfillment of a goal in the external world (agents perceive their action execution) reduces the quota of affect of the corresponding drives. For instance, if the agent perceives that he shared a Schnitzel, memory search informs which drives this action is associated to and hence reduce their quota of affect.

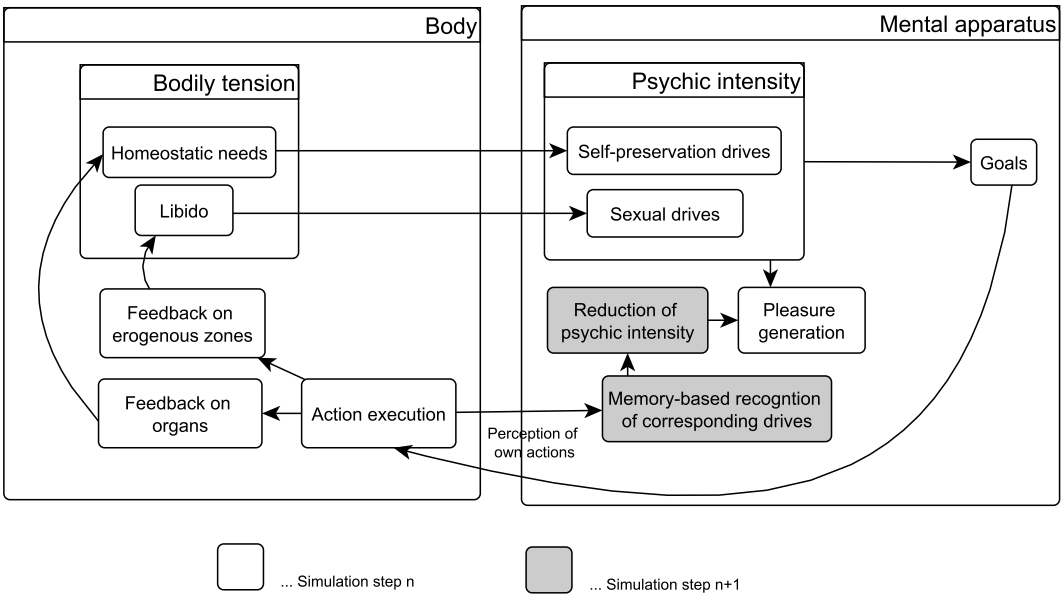

Figure 4.5: Drive cycle (SDD14). The success of executing drive-based goals can be recognized by bodily feedback or by perceiving how the agent executes the goal. This feed back results in according pleasure in the subsequent simulation step of the model.

4.3.2.4 Neutralized Intensity

Following psychoanalytic theory, the function of neutralization is central to functioning of a developed psyche (DBD+15, p.84). It provides psychic intensity from drives to higher cognitive functions and hence its usage for the secondary process. The general concept (described in (DBD+15, p.125) by the author of this thesis) was already conceptualized in the SiMA project, but the modeling details and implementation was missing. Additionally, the details of how it is used in the secondary process was not clear: In this thesis one example is provided and implemented by showing how neutralized intensity determines the processing modality of emotion and feelings in decision making (see Chapter 4.3.3.5).

Neutralization (function module F56 Neutralization, see Figure 4.1) works by splitting a share (the ratio is defined by personality parameters for self-preservation and sexual drives separately) of all drives' quota of affect. The resulting neutralized intensity can be used according to the principles of the secondary process (the drive's quota of affect is used according to the principles of the primary process, see 4.2.1).

Technically, a cyclic storage (DT3 in Figure 3.8) is used to temporarily store the gained neutralized intensity for the subsequent distribution to other modules. Since multiple function modules send demands of neutralized intensity, the function module F56 has to determine how to distribute the currently available neutralized intensity. If the amount of neutralized intensity does not suffice all demands, the distribution depends on the modules' priority, which is based on a modules' personality-dependent relevance of underlying mental function, the demanded amount of neutralized intensity, and the previously used neutralized intensity.

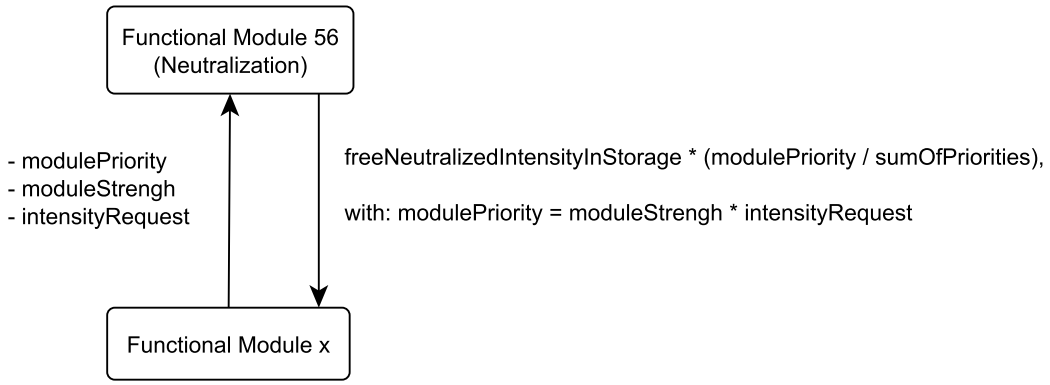

Figure 4.6: Distribution of Neutralized Intensity. Processing in the secondary process is dependent on neutralized intensity. Function modules request neutralized intensity from the neutralization module, which distribute it based on the module's current priority, its request, and its general strength.

Finally F56 generates pleasure according to the amount of currently used neutralized intensity, since it is a share of psychic intensity.

As mentioned, a concrete example of the impact of neutralized intensity is shown by its usage for regulating the impact of emotions and feelings in decision making (see Chapter 4.3.3.5).

4.3.2.5 Summarizing Exemplification

The low fill level of Adam's stomach and his low blood sugar level are send as body signals to the mental apparatus, where they get symbolized into quota of affect. Since Adam is an impulsive personality (personality parameter), the body signals are mapped into a high quota of affect. Together with the information about its source, the quota of affect is represented as a drive candidate. Due to Adam's personality, this drive candidate is split into an aggressive drive component with a high share of the quota of affect, and a libidinous one with a lower share (the split factor follows a personality parameter). Both drive components search for ways to satisfy the drive source by searching for memorized objects and actions that are associated with satisfying this drive. The number of activated memories is dependent on the drive component's quota of affect. The found memorized object-action pairs are valuated based an their memorized

ability to satisfy hunger and the hunger drive's current quota of affect. Since eating a Schnitzel is memorized as the best possibility to satisfy hunger, it is selected as a drive object and drive aim. With this information the drive component become a drive representation. However, all other activated object-action pairs are also associated (with lower association weight) to the hunger drive, and may be switched in further processing (e.g. in case a Schnitzel is not available). The same memorized object and/or action may also get activation from other sources. Since the Schnitzel has also satisfied the sexual oral drive, it gets additional valuation. External perception of similar objects and/or association to already activated objects may activate the Schnitzel further. The activation of all these sources is aggregated under consideration of the current relevance of the source. For instance, since Adam's drives are currently very high, the Schnitzel gets high activation and hence valuation from drives. And eating the Schnitzel is highly priorized, even if it does not get activation from the external environment. After eating the Schnitzel, Adam's stomach's fill level and blood sugar level rise. The mental recognition of the drive's resulting changing quota of affect lead to generation of pleasure.

4.3.3 Emotion

The drive concept provides a basis for motivation and decision making in autonomous agents. However, as analyzed with the exemplary case, the requirements of a simple agent interaction demonstrate the need to extend the drive concept with a concept of emotion. Summarized, drives enables goal-directed valuation, but only emotion represents the agent's internal state holistically for valuating goals - by considering unpleasure and pleasure in integrating the contextual internal and external situation - and expresses the state to the outside world, which enables adaptiion on other agents' state.

4.3.3.1 Approach

Emotion is modeled as a holistic (meta)representation of an agent's internal state, integrating different sources to indicate the agent's internal state (to the mental apparatus and to the outside world): bodily needs, (body) perception, norms, and imagination. By integrating *state indicators* from different demands and affordances, emotion mediate between them and links the internal and external world. Hence, emotion integrates information about the bodily and psychic state holistically and express the state externally. Thus it can be regarded as a holistic psycho-somatic state indicator that gives an absolute indication (additional to the indication relative to a demand, e.g. by drives) about the agent's state. Therefore emotion is able to activate and valuate data that is able to account for an agent's current state in the current external world. The perception of the expressed emotional state can be regarded as feeling, which can be used to reason about the sources of this state.

Following the general approach of using state indicators as means for valuation for the sake of enhancing the state, emotion can be regarded as a holistic valuation mechanism that provides a comprehensive means to decide an agent's behavior. By integrating (possibly conflicting) valuations and including the external state, it integrates the Id, Super-Ego, and external world affordances. Hence, emotion is a comprehensible and flexible mechanism to carry and integrate multiple internal and external influences on decision making.

However, emotion as a holistic representation integrating various sources of bodily and psychic state indication does not replace these sources, but only provides a meta-representation of them

for the sake of integration and contextual consideration. In this regard emotion can be described as a summarized state representation that is used as a summary valuation - additionally to the directed valuation of the different sources, e.g. drives.

In integrating the external and internal world, also a social perspective is considered in the concept of emotion. With other agents being a part of the environment, emotion is used to valuate them and adapt on their expressed emotional state.

4.3.3.2 Emotion Formation

Emotion in the SiMA model is represented by a data structure that is formed in a two-step process. First a vector integrating the sources of emotion is formed based on the agent's actual drives and perception. The resulting vector of *basic emotion* can be ascribed with a subset of following labels: anger, mourning, anxiety, joy, saturation, and elation. Therefore the function module F63 in the SiMA model was introduced. Basic emotion, then, can be transformed into extended emotion by defense mechanisms. That is, the emotion vector is extended and can be assigned additional labels (e.g., guilt, pride, envy).

Basic Emotion

The basic holistic state of an agent is represented by following *emotion factors*: (1) quota of affect from drives, (2) pleasure from drive satisfaction, (3) former state indicators associated to memories triggered by perception or imagination (i.e., indirectly activated memories). Hence, emotion as the corresponding state indicator can be represented as a vector of following emotion factors: pleasure, unpleasure, quota of affect from libidinous drives, and quota of affect from aggressive drives (see Figure 4.7).

The bodily aspect of emotion as state representation is sourced in bodily needs (i.e. drives) and body perception. However, in both cases their psychic representation is used. In the case of bodily needs, this is the drive representations' quota of affect, which adds to the unpleasure in the emotion vector and to the libidinous and aggressive factor. In case of body perception, this is the body representation's associated emotion valuation: As with every incoming sensory data, bodily sensory data are matched against memories, which are associated with valuations of pleasure and unpleasure. In case of the body, sensor values from proprioception and from external perception are matched - mediated by a memorized body schema of oneself - against bodily features of body states that indicate their pleasure and unpleasure valuations.

The psychic aspect of emotion as state indicator is sourced in memorized valuations activated by external perception and imagination. They are activated by perceived situations. Using spreading of activation (WGF+15) in the agent's associated memories, memorized situations that are activated by the agent's perception subsequently activate memorized emotions, which impact the agent's actual emotions via their four emotion factors. In the course of spreading activation the association strength of memories with emotions co-determines their activation. Hence, the stronger a memory is associated with an emotion, the more likely it will be activated and influence the agent's actual emotional state. This boost emphasizes the impact of emotions in a memory-based agent.

As described in Chapter 4.3.2.3, pleasure is gained from bodily and psychic feedback, which feeds the pleasure factor of the emotion. It must be emphasized that the pleasure and unpleasure factors are orthogonal. That is, the pleasure indicated in the emotion vector is not dependent on

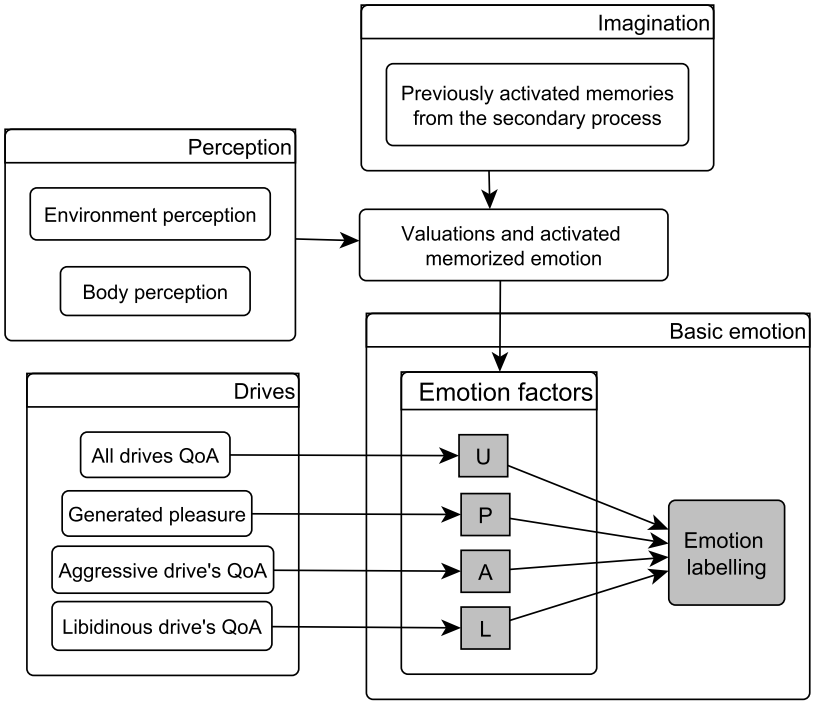

Figure 4.7: Formation of Basic Emotion in the SiMA Model (SDW+13). The vector for basic emotion integrates the bodily state represented by drives, and the mental state represented by valuated memories. The resulting vector can be labeled with folk psychology terms, such as anger and anxiety in case of dominant aggressive unpleasure

the unpleasure factor in the vector. Moreover, emotion indicates the currently available pleasure and unpleasure.

The impact of the different sources (drives, perception, imagination) on the emotion factors is personality-dependent and hence is regulated by a personality parameter. But besides this explicit regulation an implicit regulation occurs: the more emotional memories are activated by perception, the more impact perception has on emotion formation.

Extended Emotion

In further processing basic emotion, i.e. the emotion vector and its labels, may be extended and accentuated by additional factors. This happens primarily by defending basic emotion (e.g. triggered by a super-ego rule 'Do not be angry') and corresponding valuations of psychic content. The result is an extension of the emotion vector by an specific factor that cannot be generalized. Examples are extension of the emotion vector that can be labeled as pride, guilt, and envy (see Chapter 4.3.3.3).

An example of extending the emotion vector is a conflict factor. In the case of a conflict between high drive wish (e.g. to eat a Schnitzel) or an emotional state (e.g. anger) and a super-ego rule that forbids the emotion, the emotion vector is extended by a conflict factor. This constellation is labeled with guilt. If this conflict is due to a super-ego rule that forbids high valuation (this can be

described as the rule 'Do not be greedy'), and the agent uses the defense mechanisms projection[18], the emotion vector is extended by a factor informing about the chosen defense mechanism. In the described example, the resulting constellation is labeled as envy.

4.3.3.3 Emotion Labeling

As described, emotion is a holistic integrative state indicator that can be represented with a vector, i.e., in a dimensional form. Hence, the singular, emotion, is used. However, different constellations of the emotion vector can be described as dedicated states, represented by labels.

The mapping from constellations of the emotion vector to word labels was done with the expertise of psychoanalysts [19] in the SiMA project. The approach was to start with the given representation of the agent's internal state, i.e., quota of affect and pleasure, and analyze which emotional labels fit best to the possible constellation of the given factors.

In particular, any dominance of the unpleasure factor (relative to pleasure) can be labeled as anxiety; an additional dominance of the aggressive factor (relative to the libidinous factor) can be described as anger; and as mourning in the opposite case. Joy is a dominance of pleasure; which can be further described as saturation in case of a dominant libidinous factor, or elation in case of a dominant aggressive factor. If no dominant relation between the factors occur, the constellation can be described by the respective labels in an aliquot manner. Hence, the emotion vector at any given time can be attributed to minimally two of these descriptions.

Given that the sources of these labels - the four factors of basic emotion - are injunctive factors of the mental apparatus, these constellations occur in every human. This is not the case for extended emotion, since it is dependent on a cultural aspect of the mind - the norms (super-ego rules) and their mediation (defense mechanisms), which may differ dependent on the personality. However, in both cases, basic and extended emotion, the selection of labels is subject of folk psychology, i.e., attributing words to (expressed) emotional states of other humans.

Different approaches aim to map these labels to underlying neural systems. However, for a functional psychic representation, a dimensional representation is sufficient and more appropriate. For most functional purposes (e.g. valuation) only the emotion factors are relevant, i.e., the causes of emotion. Most importantly, the agent is not able to reflect on these emotion labels in the primary process. Beginning with defense mechanisms, also a recognition of a emotion constellation is functionally relevant (e.g., with the norm 'Do not be angry').

4.3.3.4 Feelings

Regarding the difference between emotion and feelings we follow Damasio's separation (Dam03, pp. 83 ff.). Emotion are objectively perceivable, i.e., they are expressed bodily and perceivable publicly; feelings are only subjectively perceivable, i.e., they are conscious perception and (verbally) interpretation of this bodily state. In the SiMA model this separation is reflected by using the concept of emotion for the primary process and feelings in the secondary process. Emotion is expressed bodily (via an data interface I5.20 between the mental apparatus and the de-symbolization function modules in the SiMA neurosymbolization layer, which translate the

[18] ... which attributes the forbidden corresponding emotion (high unpleasure and pleasure) to a currently available agent to whom a positive relation (i.e. a memorized association and pleasureful valuation) exist.

[19] Klaus Doblhammer and Zsofia Kovacs

mental representation of emotion into body expressions, see also Fig. 4.9 and Chapter 4.3.3.6) and perceived as feelings. However, since a detailed generation of feelings is not a focus in this thesis, a shortcut is chosen by directly transforming emotion into feelings (see transformation track in Figure 3.8) based on their intensity. The major difference is the association of emotion to word presentations and their access via them. In this regard feelings are (pre)conscious emotion, which enable the agent to reflect on its internal state and how possible goals may change it in decision making. Regarding terminology, different to emotion the plural feelings is used, since the agent assigns word presentations - again, labels - to access his internal state (represented as emotion vector), which we term feelings.

4.3.3.5 Emotion and Feelings in Decision Making

Until now emotion formation and emotion labeling was described. But how does emotion influence decision making? The impact can be summarized as emotion activating memories and feelings valuating or evaluating them[20] - depending on the available neutralized intensity (see below). Activating the agent's memories helps to determine what the agent did in a similar emotional situation, and enables reflecting about how a goal may impact the agent's state. This requires the determination of how feelings are stored in the agent's memory, its distinction to emotions, its activation, i.e., matching of the current feelings with memorized feelings, the evaluation of matched plans by feelings, and the integration of the matching process into goal selection in decision making.

Emotion and feelings operate in decision making by activating and valuating memories. Hence, first we have to describe how emotion and feeling are memorized. Overall, memories are anchored with emotion to valuate them. In a technical sense, this enables using current emotion as a extended search pattern[21] to enhance activating appropriate memories, by considering the agent's emotional state. This happens in the SiMA functional module F46 Memory Traces for Perception (see Fig. 3.8), where all images similar to the current external (environmental perception) and internal (drives and emotion) situation are activated. In particular, the better the current emotion matches to an emotion associated to a memorized image, the more this images gets activated. Therefore the spreading activation framework presented in (WSG+13) is used and extended. Given sequential processing of the functional modules in the SiMA model, emotion have to be feed back from secondary process to F46 (Interface I5.19, see Figure 4.9).

To store emotion in the agent's memory, in principal the same data structure as describe above (see Figure 4.7) is used. Following a principle of the SiMA approach, in the primary process (opposed to the secondary process) data cannot be accessed by their name, but only by their defining properties. In case of emotion these are the mentioned emotion factors (pleasure, unpleasure with its libidinous and aggressive parts). Feelings, on the other hand, can be accessed by their word presentations, i.e., labels such as 'Joy'. Hence, the stored data for emotion and their according feelings are the same. However, emotion are fetched using the emotion vector as a search pattern, feelings are fetched by using the label as a search pattern. That is, emotion and feelings are only different access ways on the same data.

[20] As mentioned, the term feeling is used to indicate another usage of the emotion vector. That is, the vector is the same, but he it is used for valuating goals differ in the primary process (usage as emotion) and the secondary process (usage as feelings).

[21] That is, the emotion vector with its factors are used additionally to other search patterns, such as perceptual features and drives (see Chapter 4.3.2.2), to activate images that fit both the internal and external situation.

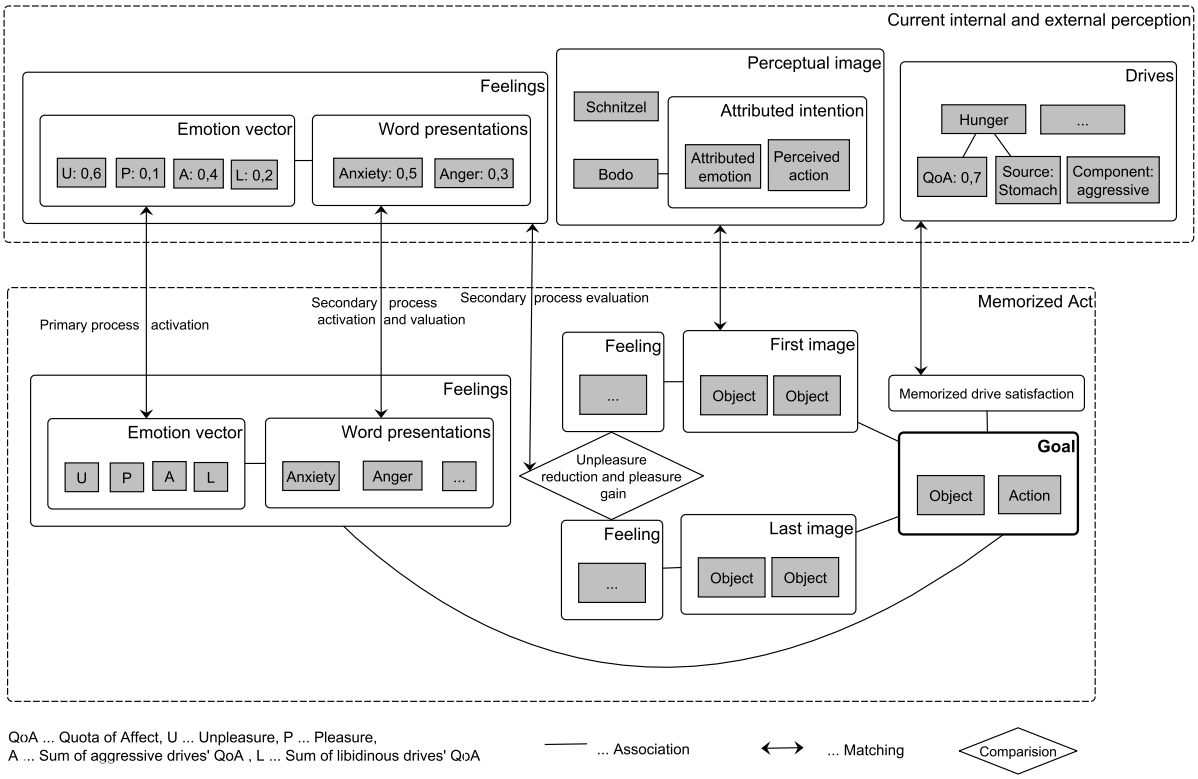

Figure 4.8: Memorized goals are activated by perception, drives, emotion, and feelings. Activation is conducted via matching and entails (e)valuation. In the primary process this happens by calculating the difference of matching pattern, in the secondary process matching additionally is conducted via evaluation, where the expected pleasure gain and unpleasure reduction (between the first image - i.e., the feelings associated to the currently perceived situation, and the last image -i.e., the feelings associated to the expected situation - of an act) is related to the agent's current feelings. The different matching factors are weighted and aggregated to determine a goal's overall adaptivity to all internal and external factors.

After activating appropriate images for the current internal and external situation in the SiMA functional module F46, their memorized feelings valuation are compared to the current feeling to valuate them. This valuation is processed by different functions in the secondary process (F26: Decision making, see Fig. 3.8), depending on the available neutralized intensity. In case of low neutralized intensity reactive valuation is done, by using the memorized emotion. In case of high neutralized intensity the agent deliberate about how the activated images would change current feelings (see below).

Valuate Goals by Triggered Emotion – Heuristic Matching

This form of goal-valuation by emotion corresponds to an impulsive reaction to a situation. In the end, it corresponds to a triggering of a goal by the agent's current feelings due to memorized associations of the goal with similar feelings. Therefore a goal's associated feelings (from memory) is matched with the agent's current feelings (see Figure 4.8). The matching value is used to set the priority of the goal according the agent's state of feeling. Overall this type of mechanism is similar to activating and valuating memories by drives, where the search pattern is given by the agent's current feelings, which are compared to memorized feelings (associated to memorized possible goals) to get a ground of what the agent did in similarly felt situations. Since emotion

and its usage as feelings is a holistic state indicator, it is able to consider the agent's abstract internal state and bridge it with the external world - additionally to drives.

Evaluate Goals by Expected Feelings – Reflective Reasoning

In case of high neutralized intensity are reflective mode of reasoning is chosen, where goals are evaluated according to their expected gain for the agent's state. In particular, every goal is associated with a plan to fulfill the goal, so-called acts in the SiMA model (see Chapter 3.2.6.2). Every act consist of a sequence of images. And, as mentioned, every image can be associated with feelings. The first image can be regarded as a precondition-situation to trigger the execution of the act. The last image of an act can be regarded as the expected resulting situation, i.e., the result of executing the sequence of actions an act involves. Given that feelings indicate pleasure and unpleasure, the agent evaluates the plan by comparing the expected reduction of unpleasure with the unpleasure of the precondition-situation, and, pleasure respectively. The accumulation of these two factors indicate the priority of the goals according to the agent's state of feeling. The relevance of the single factors is weighted by the agent's current unpleasure and pleasure. That is, if the agent currently has high unpleasure, the relevance of unpleasure-reduction has a higher impact on the evaluation of the plan.

Integrate Feeling-Evaluation into Multi-level Evaluation

As mentioned, goal-evaluation by feelings is the last valuation-mechanism in an incremental multi-level procedure to determine which goal and plan to choose. Hence, an integration of the priority-value given by the described goal-evaluation by feelings into the priority-values from other valuation-mechanisms (e.g. drives) is required. In this regard it may be possible that a goal has no priority-value from other valuation-mechanisms. This is the case in the fleeing-scenario (see Chapter 6), since drives do not trigger a goal to flee (it does not fulfill any drive aim). The integration is processed in a goal-class' method. To consider the global normalization scale of 0-1 and the comparability of the goals' total priority-value, a scaled accumulative function is chosen:

$$w_{i+1} = w_i + (1 - w_i) * q_{i+1} \tag{4.1}$$

, where w is the accumulated priority-value, i is the valuation-step, and q is the current valuation by the feeling's priority-value.

4.3.3.6 Emotion Expression

An effect of emotion is its bodily expression. This includes changes in organs and their expression, by which the agent's internal state is expressed externally, e.g. enabling other agents to adapt their behavior on that state (see Chapter 4.3.3.8).

The bodily expression of emotion enables subsequently their perception as feelings. Hence, emotion (and not feelings) are bodily expressed. For bodily expression of emotion in the SiMA model the function modules 'F67: Bodily emotion reaction', 'F68 Actuators for glands' (with a renaming of 'F32: Actuators for muscles'), and an interface from 'F71: Composition of extended emotion' to F67 is developed (see Fig 4.9).

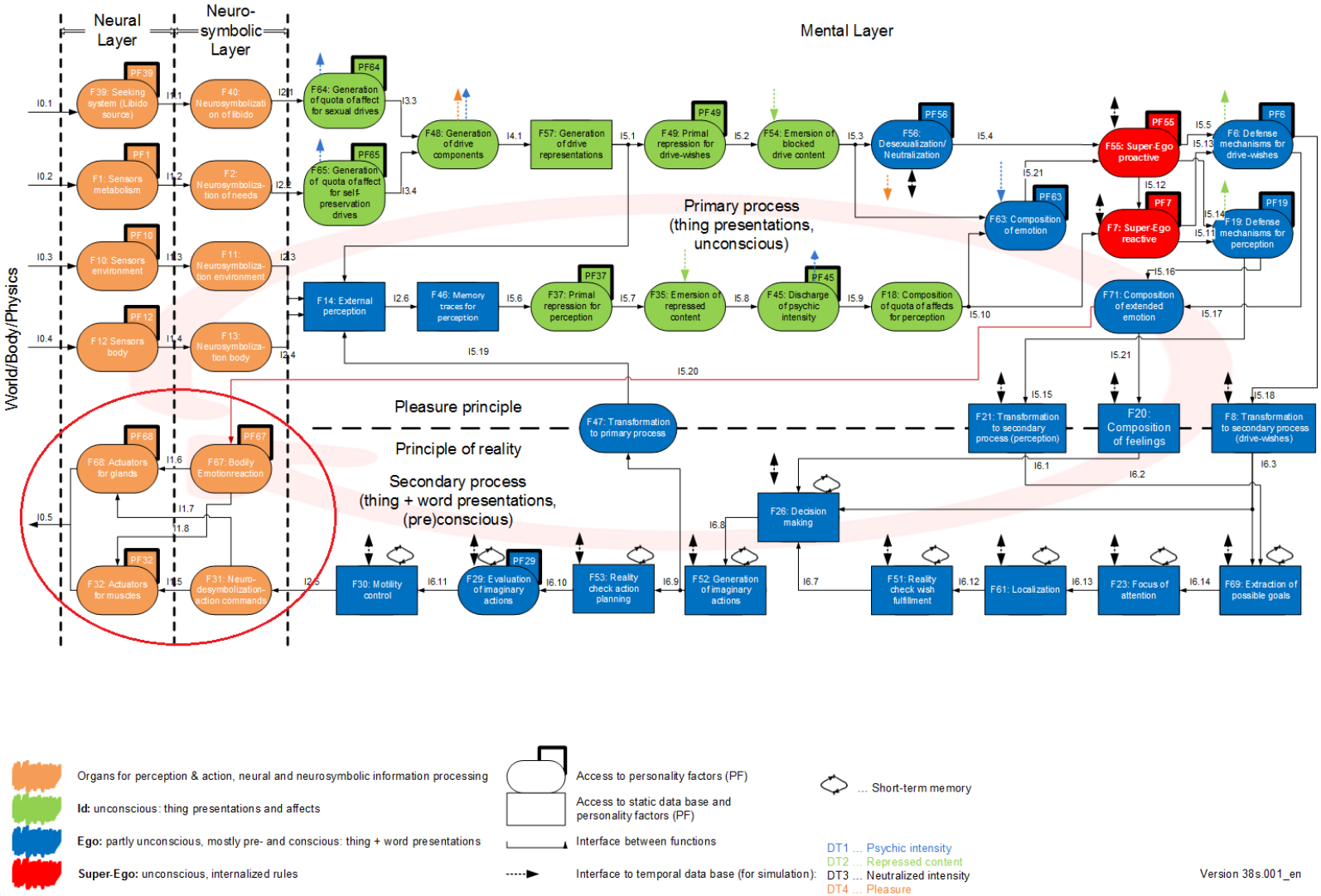

Figure 4.9: Emotion are expressed in the body via interface (I5.20, highlighted in red) to de-symobilzation functions in the neurosymbolization layer (encircled in red).

The main purpose of these function modules is to determine the causal chain from emotion to organ variables and subsequently to expression variables (including visualizations in simulation). Therefore the effect of emotion on expression variables via organ variables is modeled. Being a functional model the focus was not a plain mapping, but the generation of the body effects following a causal chain.

The considered body variables and expression variables are shown in Figure 4.10. However, the goal was no exact reproduction of the human system and relation between the variables, but only a demonstration of the principle of emotional bodily expression and its impact when interpretated by other agents (see Chapter 4.3.3.8). Registering and signaling the change of emotion (and hence regulating the impact), and the connection to body variables (body parts and organs) and expression variables are implemented with the concept of internal action executors (IAE), representing control by hormones or muscles. IAE executes an internal action command (IAC), e.g., to increase a specific body variable. IAC execution can be parametrized to account for the different impact an emotion has on an body variable (e.g. anger increases the heart rate more than anxiety).

Additionally to the mentioned changes in the SiMA mental apparatus, the extension of the agent's body was required. This includes the introduction of the corresponding body and expression systems (extending clsComplexBody with clsBodyOrganSystem and clsFacialExpression including classes for every body and expression variable) and their connection via IAEs and IACs. The expressed variables are then visualized in simulation (using overlays, e.g. see Figure 4.11) and attached as physical attributes of the external world to enable their perception by other agents.

4.3.3.7 Summarizing Exemplification

Adam's current internal state is caused by his current drives and the memories activated by perception. He is anxious and angry due to high aggressive unpleasure coming from his high hunger, and due to memories of anger and anxiety activated by perceiving Bodo. His super-ego (internalized social norms) is triggered by his high anger, which is forbidden. This results in an inner conflict. Due to sufficient neutralized intensity, which is split from the drives' quota of affect (the ratio is determined by a personality parameter), Adam's defense mechanisms are able to deal with the conflict by gradually shifting value from the aggressive emotional factor to libidinous emotional factor. The resulting adapted emotion vector is labeled as anxiety, mourning, and guilt, which Adam has learned to ascribe to this felt internal state. In this form Adam is able to access his emotional state as feelings in the secondary process, where data is able to get conscious. In the primary process, drives have activated associated memorized objects and actions, and emotion has activated associated memorized images. Using the ascribed word presentations (i.e., labels), Adam has enough neutralized intensity (as a measure of processing in the secondary process) to uses his feelings (as the representation of his internal state in the secondary process) for reflecting on the expected benefit of the activated goals. That is, Adam relate the expected pleasure gain and unpleasure reduction to his current unpleasure and pleasure. Since his current unpleasure is much higher than his pleasure, unpleasure reduction is more important. The goal to share the Schnitzel fulfills these criteria. That is, guilt activates the goal and together with a reflection on the expected unpleasure reduction, it his higher ranked than other goals.

Adam's emotion (i.e., the emotion vector in the primary process) is expressed bodily. Adam cannot control how his internal state is expressed externally (this is only possible afterward - via feelings as the consciously felt internal state).

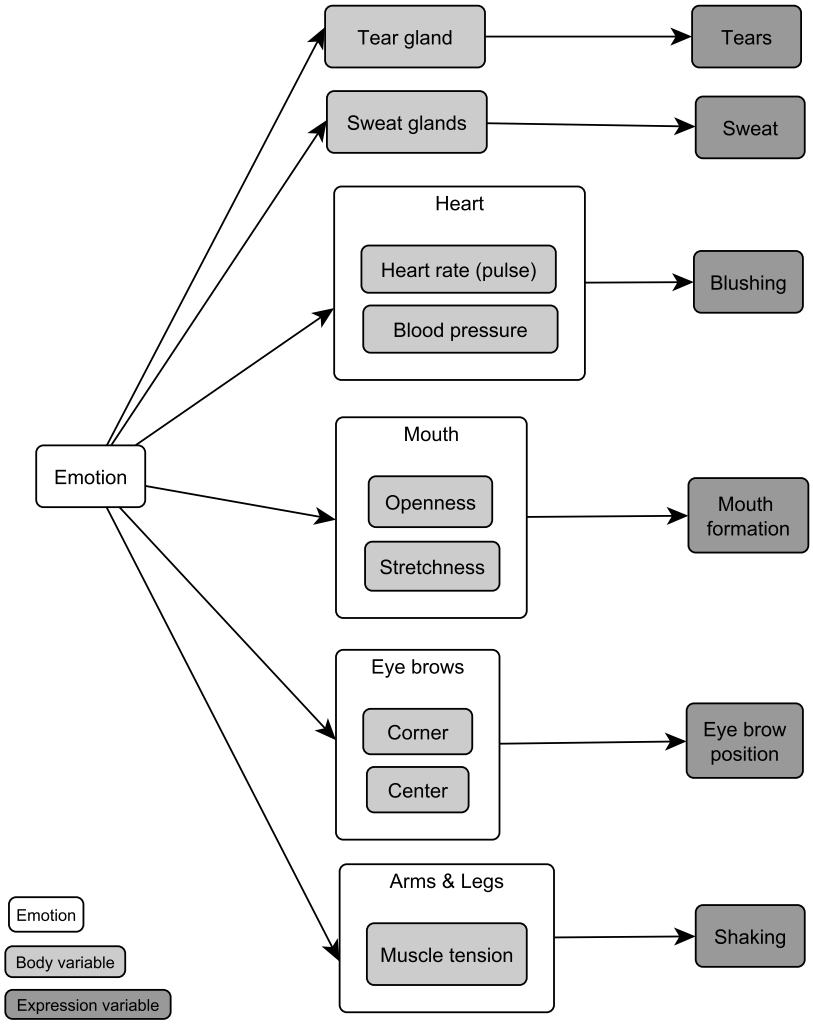

Figure 4.10: Bodily Expression of Emotion. Emotion causally influence body variables that are expressed externally via expression variables. The relation (n:m) between the different variables are configurable by the concept of internal action executors and internal action commands.

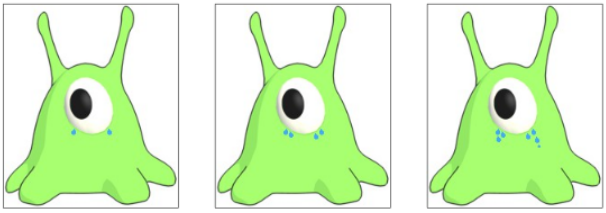

Figure 4.11: Crying agent in simulations, as an example of bodily expression of high mourning.

4.3.3.8 Emotion in Social Context

As a result of bodily expressions of emotion, other agents may interpret the agent's internal state. Thereby emotion are attributed, which in turn affects the agent's own emotion (and impacts decision making implicitly) and enables to interpret the other agent's intentional action. Both

capabilities enable social adaption and can be regarded as mental functions for social cognition.

Emotion Attribution

After recognizing a percept as an agent, the perceived agent's bodily expressions are recognized as emotion. As with any other perception this consists of activating similar attributes in the perceiving agent's memory. In this case, perception activates similar body states (defined by body expressions described in Chapter 4.3.3.6), which are memories associated to a memorized body schema (see Figure 4.12). In particular, a body schema is a memorized representation (Thing presentation mesh - TPM, see Chapter 3.2.6.3) of an agent's perceived shape and a body state (also TPM) as associated information about the emotional state for a specific body state (i.e., the unpleasure and pleasure state for an perceived bodily state).

In case of a perfect match of perceived and matched body state, the memorized body state's associated emotion is also associated to the perceived body state, otherwise graded category membership is conducted as sketched in Chapter 4.3.2.2 and the resulting emotions are weighted and merged into a single emotion vector, which is then associated to the perceived body state.

In most cases the perceiving agent uses information of its own body states. Only if the agent has enough experience with another agent and if it differs from his, bodily states and associated emotion are explicitly stored for another agent (e.g. for agent x blushing is not a sign for anxiety as with me, but a regular bodily feature).

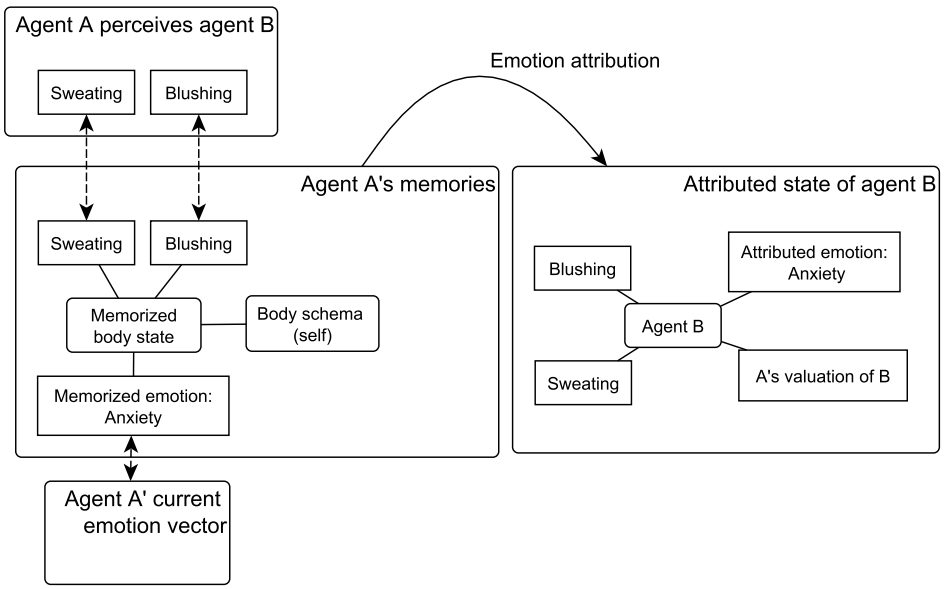

Figure 4.12: Emotion attribution and transfer.

Emotion attribution is implemented as an extension of perceptual categorization in the SiMA model (see Chapter 3.2.6.2). As with other memories, body state memories may be activated by perceptual stimuli (given by the perceived bodily expressions) or by the agent's internal state (given by current emotion). The latter can be regarded as affective priming, in particular as modulating emotion attribution by mood congruent perception (SM10): the current own emotion supports the recognition of similar emotion in others and impair recognition of opposed emotion.

Besides influencing the attributed emotion by the perceiving agent's own emotional state, it is influenced by the valuation of the perceived agent. The recognition of emotion in other agents

(via their expressed bodily state) implicitly recognizes the perceived other as agent (besides an object). In this regard the other agent is not only valuated regarding its ability to satisfy drives (see integrated drive object categorization in Chapter 4.3.2.2), but also valuated in a holistic general sense with an emotion. Therefore, as usual, memorized emotional valuations of similar agents (i.e., exemplars, see Chapter 4.3.2.2) are used. As in integrated drive object categorization, exemplars are activated by similarity and by expectation, which in this case is activation by the perceiving agent emotional state. The resulting emotional valuation in turn is also influenced by the perceiving agent's own emotional state.

The emotional valuation of the agent is an emotion vector, consisting of a mean value from all associated emotion vectors of k similar memorized agents, weighted by their activation values.

Overall, the result of emotion attribution is the association of a body state (generated based on memories) including an attributed emotion. Since emotion attribution happens in the primary process, it is unconscious data and subsequently impacts the agent's (unconscious) own emotion.

Emotion Transfer

After emotion is attributed to a perceived agent via an associated body state, it affects the agent's own emotion (in the primary process). As previously described, memorized emotion activated by perceptions influence the formation of current emotion (see Figure 4.7). The same principle is valid in the case of perceived body states. However, additional to a direct impact of attributed emotion to an agent's current emotion (i.e. homologous impact on the same emotion factors, which corresponds to the phenomena of emotion contagion), impact occurs based on the relation and valuation of the perceived agent. For instance, based on the valuation of the perceived agent, the attributed emotion may have an heterologous impact on the perceiving agent's current emotion (i.e., attributed unpleasure affects pleasure, phenomenologically described as malicious glee or Schadenfreude). The ratio of these two types of impact by attributed emotion is regulated by a personality parameter.

To account for both kind of emotion transfer, the values are aggregated for each emotion factor. Formula 4.3.3.8 exemplifies it for pleasure.

$$currentPleasure = currentPleasure + +transferedPleasure$$

transferedPleasure = contagedPleasure ++ valuatedTransfer

valuatedTransfer = [aP * (vP / (vP + vUp))] ++ [aUp * (vUp / (vP + vUp))]

Where x ++ y symbolizes 'non-proportional aggregation': x = x + (1-x) * y
(4.2)

aP ... attributed pleasure
aUp ... attributed unpleasure
vP ... valuated pleasure
vUp ... valuated unpleasure

In case of emotion contagion the impact of attributed emotion is not dependent on the perceived agent, but is generally regulated by a personality parameter. That is, how much the agent's

current emotion get influenced by another agent's emotion is independent from the (valuation of the) other agent.

Emotion transfer of an attributed emotion based on the valuation of the perceived agent can be described as affective empathy. However, since a functional description is aimed, the functional terms of emotion attribution and emotion transfer are used, which may explain the basic process behind affective empathy.

Due to emotion transfer and the impact of emotion and feelings on decision making (see Chapter 4.3.3.5), attributed emotion has an implicit impact on the agent's decision. By using attributed emotion to deduce the intention of a perceived agent additionally it has an explicit impact.

Intention Attribution

The attributed emotion and the perceived action of Bodo are used as a search pattern to recognize Bodo's intention based on Adam's memories. This attributed intention (i.e., the attributed emotion and the perceived action) is used as an additional activation pattern to activate suitable memorized images that provides a basis of which goal to execute in the current internal and external situation (see Fig. 4.8).

4.3.3.9 Summarized Exemplification

Adam perceives Bodo and valuates him as a drive object (for reducing phallic sexual tensions), but also as an agent. Recognizing bodily expressions similiar to his, lead to considering Bodo as an agent. Using his own memorized associations about which emotion he has in a similar bodily state, Adam attributes an emotional state to Bodo. Since he associates sweating and frowning with anger and anxiety, he attributes this emotion to Bodo. This attribution is also influenced by Adam's own current emotional state. His anger impacts the attributed emotion negatively. This is also the case for the emotional valuation of Bodo, which serves as a summary valuation for all experiences with him or similar agents. Bodo is memorized with anger. The attributed emotion in turn affect Adam's own emotion in two ways. First, the attributed emotion affect his emotion directly - independent from who the other agent is - which increases Adam's anger. Second, based on the negative valuation of Bodo and Adam's personality, Adam's attributed unpleasureful emotion increases Adam's pleasureful emotion factor (which corresponds to Schadenfreude). This unconscious emotion attribution and emotion transfer in the primary process is the basis for intention attribution in the secondary process. Adam thinks about which goal he would intend when being in a similar emotional state and when executing the actions he currently perceives. Since Bodo is attributed to be angry and anxious and since he is perceived to move towards the Schnitzel, Adam attributes the intention to eat the Schnitzel to Bodo. This attributed intention in turn activates (together with the other mathing factors, i.e., guilt, drives, external perception) the goal to give the Schnitzel to Bodo.

4.4 Summarized Integration of Specified Concepts

Following an evolutionary approach (cf. Chapter 4.2.2), drives, emotion, and feelings can be regarded as incremental nested layers of the agent's state indicator, used for different valuation mechanisms with the aim to lead the agent into a better state. Hence, valuation is an incremental process that is conducted with an incrementally extended state indicator that value data under

different scope and principles. Since every layer of the state indicator is based on another layer and gets additional input from the internal or external world (see Fig. 4.14), its usage in valuation enhances adaptation and bridging the different input. This corresponds to an incremental extension of holistic control mechanisms with different focus.

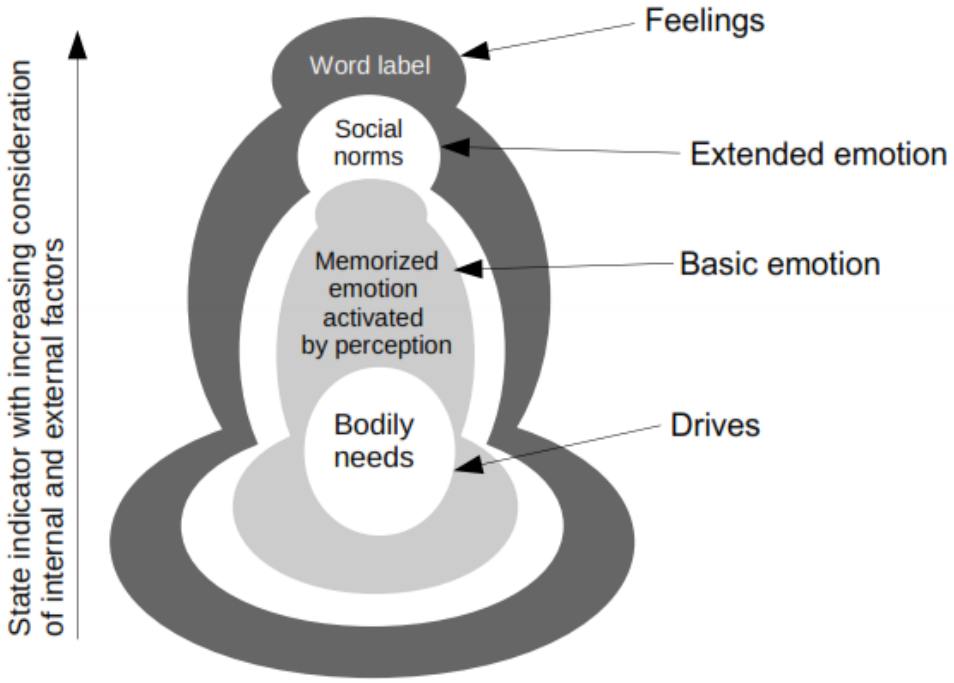

Figure 4.13: Different concepts to indicate the agent's state internally and externally. Following an evolutionary approach, different concepts are used to adapt on the agent's internal and external state, which are represented by drives, emotion, and feelings. Each of the can be regarded as a layer of the agent's state indicator with incremental input about the internal and external world. Higher layer only use certain aspects of lower layers. Each layer operates independent from the higher layers.

The hierarchical relation in the state indicator is depicted in Fig. 4.13. The basic concept for indicating the agent's state are drives, representing the basis of human control - the body and its needs. Hence, drives' quota of affect are the core layer of the state indicator. The consideration of how the agent's internal state adapts on the external world is considered in the second layer of the state indicator - basic emotion - which can be regarded as a psycho-somatic concept. Perceptions activate memories that indicate how the agent's state is expected to change when confronted with the current external world. This includes perception of other agents and consideration of their expressed internal state. With consideration of how internalized social norms (the super-ego rules) impact the agent's internal state, the concept of extended emotion emerged. The last layer of the state indicator is represented by the concept of feelings, which consider conscious possibilities to indicate the agent state, in particular language labels, which are learned translations of bodily feelings to words.

The usage of this incremental state indicator for valuation is depicted in Fig. 4.14. The most basic assignment of value to objects and actions comes regarding value for the body, i.e., using the drives' quota of affect. To consider the expected change ot the agent's state by the external

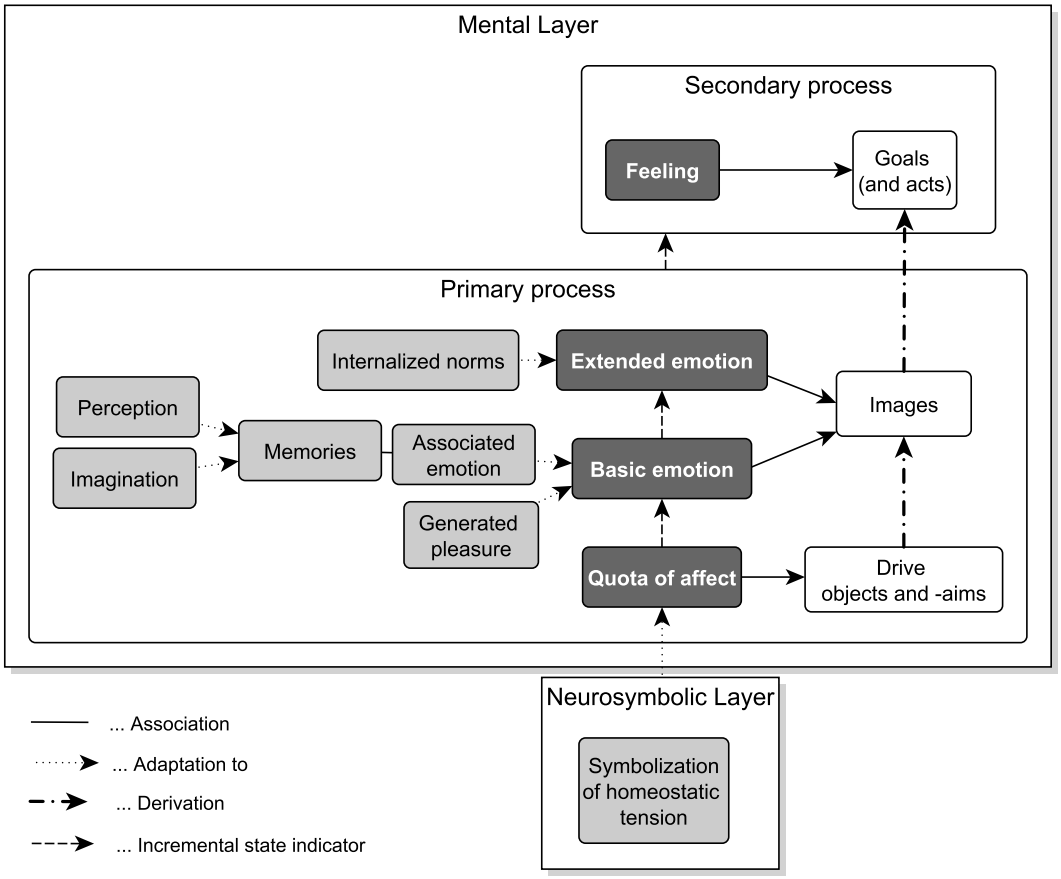

Figure 4.14: Different concepts of the agent's state indicator are used to consider different aspects for the sake of enhancing the agent's state by valuating possible goals in decision making.

world, emotion activate and value memorized images. By considering the whole situation (not only single objects as with drives) and both bodily and mental aspects it can be regarded as a holistic valuation mechanism. In this regard emotion serves a summary valuation of the whole situation - also regarding spatial and time aspects. Different to drives, which only consider pleasure maximization, emotion also consider unpleasure minimization. Valuated images include previously valuated drive objects and aims; subsequently goals are extracted from these images. Until now valuation mechanisms operate in a reactive way. Dependent on the personality and the available psychic resources (neutralized intensity), feelings may be used to reflect on the internal state and relate it to its expected change by a goal. In doing so, feelings consider the long-term implication of goals and their acts, i.e., their execution in the world.

Overall, different layers of state indicator include the lower layers but use their input about the internal state differently in valuating possible goals. It has to be emphasized that valuation is able to but do not have to operate incrementally. Dependent on the agent's state and resources and significance of a goal for the state indicators different scopes and principles, a goal may only get valuation from a subset of these state indicators. In the end only the aggregation of the different valuations is considered in selecting the most relevant goal for the current internal and external situation, bringing the best expected change of the agent's state (i.e., increasing pleasure and reducing unpleasure).

5 SiMA-C: a Foundational Mental Architecture

> *The metallurgist, knowing everything about metal molecules and their alloys, is not necessarily the best authority with regard to the functioning of car engines, even though these consist mainly of steel and iron.*
>
> Dietrich Dörner, Bauplan für eine Seele, via (Bac11, p. 62)

The research questions of the thesis at hand are tackled with two exemplary cases (see Chapter 4.1) in sequence. Although in both exemplary cases the used concepts are the same, the model and implementation requirements differ. The reasons can be summarized by (1) the different requirements of a model for an Artificial-Life simulation (simulating behavior, as in the former exemplary case) and of a decision-making simulation (only simulating decisions, as in case of this exemplary case) (2) the different requirements for a test-bed for translating concepts about the human mind into a computational representation and for applying an model for subsequent applicative questions. For basic research, we use a conceptual model in Artificial-Life simulations to translate the concepts into a computational model for the sake of testing and exploring them. Next, we can identify the functional essence of this conceptual model. This also enables an integration of the different concepts into an unified and parsimonious form (especially if the model is 'naturally grown', as the SiMA model).

Hence, for the first exemplary case a conceptual model is developed and then integrated in the SiMA model, and simulations are conducted using an implementation of it. For the second exemplary case, the SiMA-C model is developed as a compressed functional derivation of the conceptual model. The insights gained in the first exemplary case are used to develop the parsimonious SiMA-C model that follows the SiMA principles and further integrates the developed concepts. This process enabled unification on the model level and the implementation level. On the model level the descriptive concepts of 'perception, motivation, emotion, and cognition' are translated into the functions of activation, valuation, mediation, and evaluation.

5.1 Model Overview: Adaptive Decision Making

The communality of the different mentioned concepts is the basis for unifying the functionality of an artificial mind: to focus on its main objective, which is to mediate between the internal

and external world, and to fulfill the different (possibly contradicting) conditions posed by the internal and external world. Information about the agent's state, which is able to integrate the different demands and affordances from the internal and external world, serves as a the core of an unified model.

The central question a model has to answer in this regard is how to choose a goal that improves the agent's state best, in short and long-term. The key assumption of the presented model is that activation and valuation are sufficient foundational functions to choose relevant goals for the adaptation on demands and affordances from the internal and external environment, and to mediate between them.

The SiMA-C model solves the problem of goal selection based on valuated memories that are activated by conflicting internal and external sources. These sources are represented as demands and affordances. Goal selection, then, may be reflected on, using emotion as a representation of an agent's current state of pleasure, unpleasure, and conflict. Model processing is triggered by a change of bodily needs or of the external world. These changes causes data activation by two sources: Demands and affordances (see Figure 5.1). Hence, the activation process determines possibly relevant data for the current internal or external situation.

Two types of demand sources are distinguished: physical (bodily demands represented by drives and pain from body perception) and psychological (activated memorized norms). Both demand sources activate memorized goals and norms that are expected to satisfy these demands and bring pleasure. Either directly by triggered activation or indirectly by spreading of activation. if an activated goal would prevent the fulfillment of a demand or if a goal's object is expected to bring harm, it may also be memorized with expected unpleasure.

Beside demands, affordances are the second source of data activation. They activate memories that are similar to the current external world and - indirectly via spreading of activation - norms that are valid for the current external situation.

The relevance of activated goals for the current internal and external situation is determined by different valuation processes, which are triggered by the different activation sources (norm, drives, perception) and considering the goal's memorized valuation for these sources. Hence, a goal may have different valuations of expected pleasure and unpleasure regarding the different activation sources. The current relevance of the single valuation is determined by the caused activation from that source. In this regard the model distinguishes (1) normative valuation, activated by normative sources, (2) bodily valuation, activated by bodily sources, and (3) summary valuation, activated by external perception or by the agent's current emotion. The first two types of valuations correspond to an expected short-term fulfillment of bodily or normative demands, the latter uses a goal's memorized summary valuation, which provides integrative holistic information about all aspects of how a goal changed the agent's state, considering long-term expectations and context. The resulting data from the valuation process can be termed effective or affective, in an etymological sense, meaning that this data should have an impact (on the decision).

Contradicting valuations in a goal (e.g., expected pleasure for normative demands, but expected unpleasure for bodily demands) cause conflicts. Additional conflicts are caused by contradictions between demands and affordances or between contradicting goals. For example, if an internalized norm is activated, goals that are unrelated to the norm would prevent its fulfillment. Therefore, they would be marked by a conflict. Two types of conflicts are distinguished: normative and reality conflicts, with ownership conflicts being a specialization of the latter. The different kind

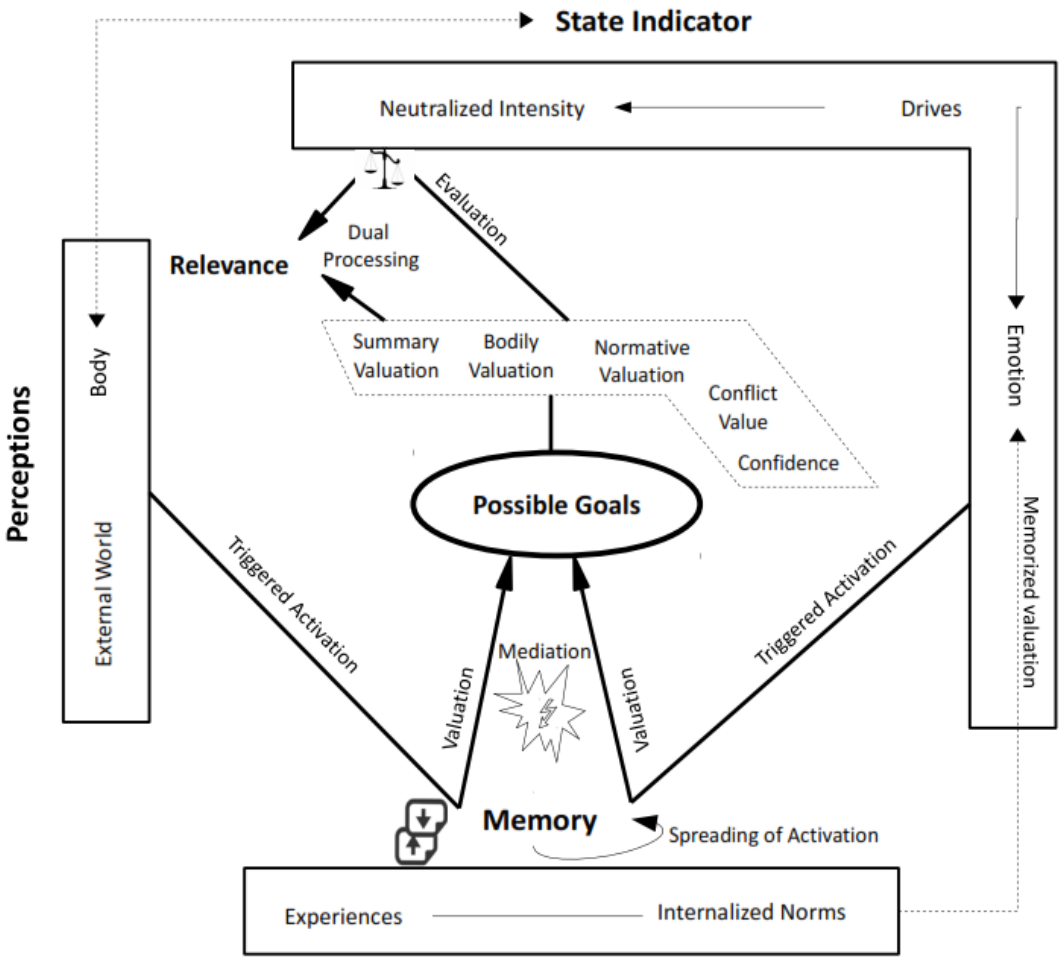

Figure 5.1: SiMA-C Model Overview. For a summarized description see Chapter 5.1.

of conflicts are addressed in processes of mediation, which operate by changing the different valuations in a goal. The result can be called arranged data.

Parallel to these processes, the generated unpleasure from the different demands and activated data, the currently experienced pleasure and conflict intensity together form the agent's state indicator, and thus represents the agent's current emotional state. A share of the unpleasure from drives is used as so-called neutralized intensity for reflective processes, regulating the grade of dual processing (i.e., between primary and secondary process).

Based on the reactive processes described so far, evaluative reasoning processes, which relate the different valuations to the state indicator, are possible. The separation between valuation and evaluation corresponds to a dual processing model of the human mind. The degree of evaluation is dependent on its necessity in case of ambiguities between goals, and the agent's neutralized intensity. The overall process corresponds to weighting the different valuations in integrating them to a single relevance value. This evaluation corresponds to a multi-criteria aggregation aiming for unpleasure minimization and pleasure maximization, while considering a goal's conflicts and the agent's confidence in the valuations. However, the guiding principle is that of a satisficing, and not an optimizing agent (cf. Simon 1956). Overall, the evaluation process results in determining

the most relevant goal for the current internal and external conditions.

Overall, the adaptive functionality of the control unit can be generated by valuation, mediation, and evaluation of activated goals. These functions are further described and exemplified in the next sections.

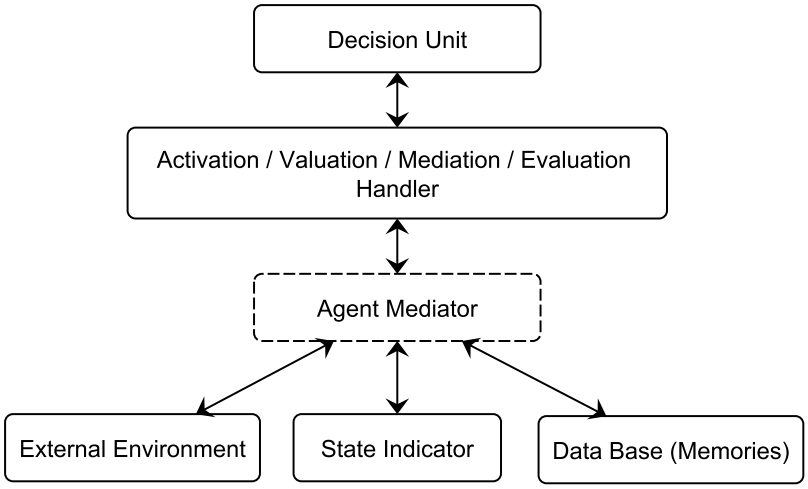

Figure 5.2: SiMA-C Software Architecture.

The software implementation should be a valid representation of the SiMA-C model. However, it follows different criteria. On the one hand, an important criterion for the model is the goal to maximize functional unification. On the other hand, the software patterns of modularization and separation of concern should be considered for the implementation. The software architecture (see Fig. 5.2) aims to consider the latter criterion, without hindering the mentioned model criterion. The decision unit module is encapsulated and communicates with the agent's external and internal world (i.e., environment and state indicator) via a mediator. This is done via 'handler' software modules. In particular, every step (activation, valuation, mediation, evaluation) of the SiMA-C model is implemented as a 'handler' modules. As mentioned, these handlers are data-driven (see next chapter) and trigger each other via data. However, handlers only communicate via the agent mediator module, which is the only one who has access to the data base and hence is able to integrate the dependencies in data-processing of all handlers.

5.2 Data Model

The SiMA-C model follows a data-driven process. That is, in a conceptual sense, the process flow is not determined by functions called with data (parameters), but by data calling functions (see Figure 5.3). Hence, before describing the model's functions, first, the data model will be described. The details about the variables in the data structures and how the data is processed is described in the respective sections.

The data model resembles the agent's mental representations, including memories. The separation between world representations and memories is only seldomly relevant (e.g. for reality check) - both use the same data structures. The key data structure in SiMA-C is the goal structure. All

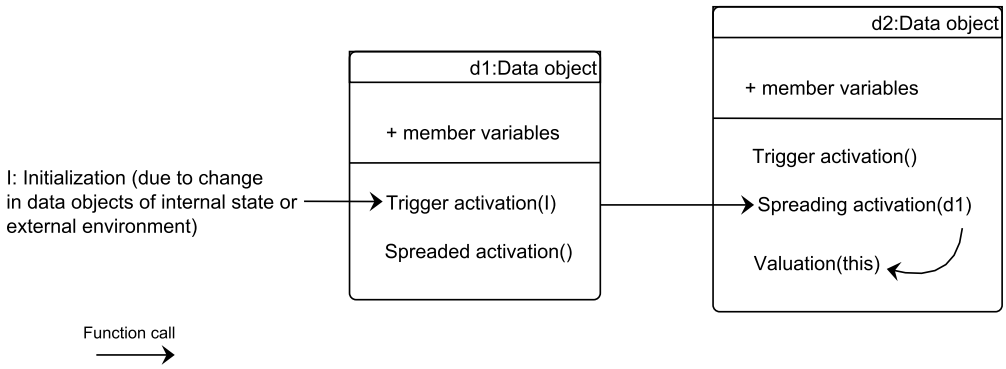

Figure 5.3: Principle of data-driven processing in the SiMA-C model UML2 and Java notion, where 'this' stands for the object itself).

processes in SiMA-C aims to determine the relevance of possible goals, i.e., activating, valuating, mediating, and evaluating them. However, goals are mainly activated indirectly via other data structures (e.g. thing presentations). All data structures that are activatable (and are able to spread the activation) consider common information about their activation. The association data type can be used for all activatable data objects, mainly between same data structures, e.g., between goals or between thing presentations. The thing presentation concept corresponds to the SiMA data structure 'TPM' and conceptually corresponds to Damasio's image concept (see Chapter 3.2.6.3). Thing presentations are mental representations of the external world (including oneself - named 'SELF') described by attributes of possibly different sensor modalities. They consider different levels of abstraction. However, for the application of the SiMA-C model in this thesis, thing presentations are only used on the object-level and situations are captured implicitly. Memorized thing presentations are associated to a summary valuation, representing how the thing presentation has changed the agent's state overall (see Chapter 5.4).

The goal structure is defined by an action on a thing presentation (e.g. to eat food). The action is executable by the agent if the action subject is the thing presentation 'SELF'. The variable 'executable' informs if the action is strategic or tactical (hence physically executable in the world). Corresponding to the different ways a goal may affect an agent, the goal structure considers different types of valuations (for details see Chapter 5.4). A valuation informs the agent about the goal's expected utility for the agent, its confidence (i.e., experience), and the conflicts associated to it (see Chapter 5.5 Conflict and Mediation). Following the dual processing approach (primary and secondary process) of determining a goal's relevance for the current internal and external situation, the variables 'valuation' and 'evaluation' relevance are distinguished.

Since normative prohibitions can be represented as commandments, norms can be represented as special types of goals, even if most norms are strategic goals (hence not physically executable) and often have abstract goal objects. However, since amongst others the impact of norms work by associations to goals (goals that are not associated to currently active norms cause conflicts, see Chapter 5.5), norms can be regarded as currently active constraints that goals need to consider. Injunctive and descriptive norms are distinguished (variable 'type') (CG04) and considered in different ways of norm activation (see Chapter 5.3).

The data structure for the state indicator is not used for the agent's memories, but only to represent all aspects of the agent's *current* state, both physical and psychological (see Chapter

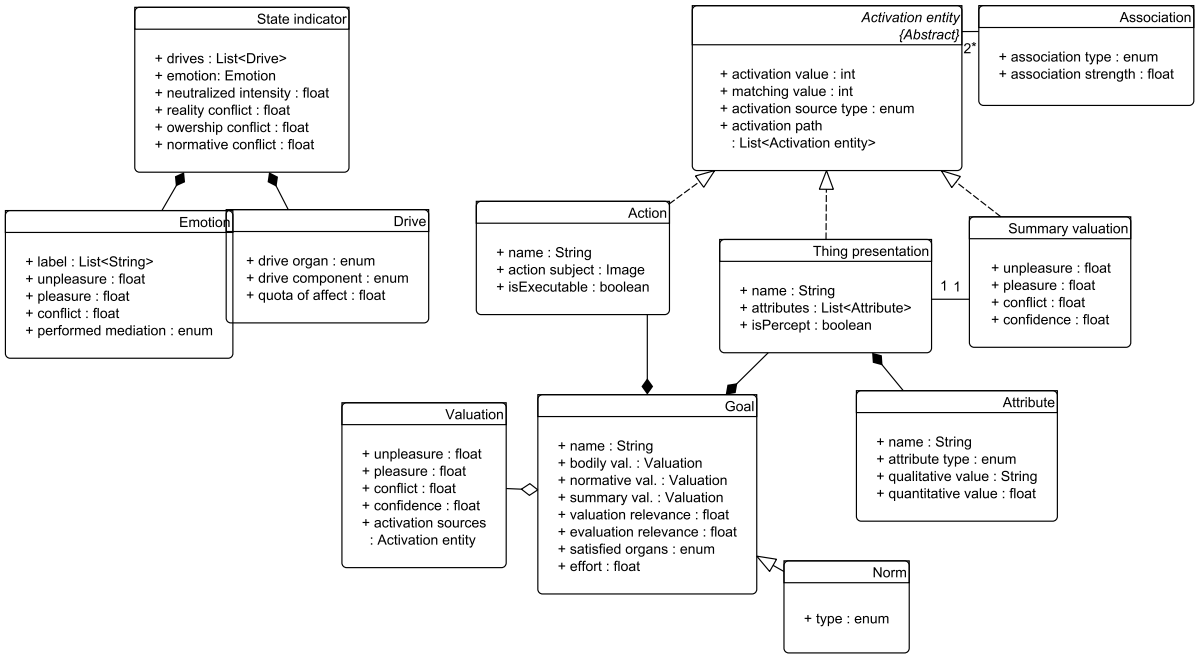

Figure 5.4: SiMA-C data model (UML2 class diagram). Adapted and extended data concepts based on SiMA data structures (see Chapter 3.2.6).

4.3.3). Hence, the concepts of drives and emotion are only used as state indicators (not in memories). As described in Chapter 4.3.2, to represent the agent's bodily needs the drive concept is used. Different to the drives' focused representation, the concept of emotion is used for a holistic summary representation of an agent's state. In particular, emotion integrate information about the agent's bodily state, (mental) conflict state, and state of activated memories (the summary valuation of activated thing presentations, see Figure 5.1). In a functional model the label of the agent's current emotional state is irrelevant, however labels are used for descriptive reasons based on the constellation of mentioned variables (see Chapter 4.3.3.3).

5.3 Activation

All memories are directly or indirectly activatable. No explicit separation of long-term and short-term memory is done in the SiMA-C model. These aspects are considered implicitly by their memorized activation value. That is, memories that have assigned an activation value from previous processing can be considered as short-termed. Data in the activation space is ordered after their current (a-priori) activation value, which they have from previous activation.

Conceptually an activation can be regarded as a requirement of adaptation: Different activation sources that require the agent to adapt on new internal and external situations activate appropriate memories. The stimuli of activation sources are percepts (thing presentations and actions), drives, and emotion. They cause a *trigger activation* process, which matches activation sources of the external (thing presentations and actions) and internal situation (drives and emotions) with memories. Next, triggered activation is spread through associated memories. In both cases, triggered or spread activation, the current activation value of data is determined by following

factors: the previous activation value, the current activation intensity, and the activation source's salience. Since usually currently activated data already have an activation value, the aggregation is given by Formula 5.1, considering normalization with a global scale of 0-1 and possibly i+1 activations of a data object.

$$a_{i+1} = a_i + (1 - a_i) * a \tag{5.1}$$

a_i .... activation value before current activation
a_{i+1} ..... activation value after current activation
a .... current activation intensity

In case of triggered activation, the activation intensity a is given by the matching value (between activation source and memorized data), weighted by the source's salience (see Formula 5.2). In case of drives as activation source, the salience is given by the quota of affect. For emotions the salience is given by their intensity (derived from their unpleasure and pleasure values). The salience of percepts of the external world are given by the agent's focus of attention, controlled by various factors, e.g., by psychophysics (e.g. loud colors) or the stimuli sources. The latter is exemplified with the case of 'Social Prompting' (see Chapter 6) by setting the message's salience based on the message-sender. That is, the perceived message about switching to green electricity is focused based on the relation to the sender-agent and the memorized valuation of the agent. For instance, if agent A memorizes a good relation to agent B, from whom A gets a message, and B is well valuated, it is more likely that A focus on the message. Such perception salience also serves as a basis for explaining social pressure and is closely with the implementation of descriptive norms in the SiMA-C model. Additionally, the SiMA-C user platform allows to configure the salience in the external world and the stored activation value in memories manually, e.g. to consider priming effects.

$$a = m * s \tag{5.2}$$

m..... current matching value between activation source and activated data
s ... salience of the activation source

The matching value between activation source and memorized data is calculated by comparing their similarity (i.e. a data base search) between the stimuli features (which identify data and are specified for each activatable data structure) and the corresponding features in memories.

Drives directly activate goals that are known to fulfill the underlying bodily need. Information about the bodily source (drive organ and drive component) is used as stimuli features of drives. The activation is determined by memorized ability to reduce the need.

External Percepts may activate thing presentations or actions (and hence, indirectly, norms). Stimuli features of a thing presentations are multi-modal attributes, and the matching value is determined by the number of similar qualitative attributes (e.g. color and shape). An action's identifying features are the action subject and action name.

Emotion activates thing presentations. Stimuli features are emotion factors of the memorized emotion that is associated with a thing presentation. Matching of these features determines the activation value.

Hence, goals and norms are only activated indirectly by spreading of activations.

After memories are trigger activated, their activation value spreads to associated memories. Models of spreading activation represent an approach to memory retrieval in a generic network architecture. Particularly, spreading activation models how activation spreads from a network-node to associated nodes. Such a process is also called *associated retrieval* (Cre97). The spreading step begins with the calculation of the incoming activation from associated nodes (Cre97) or with the initial activation, respectively. Thereby the association weights may be considered:

$$I_j = \sum_i O_i * w_{ij}, \tag{5.3}$$

where
I_j is the total input of node j;
O_i is the output of unit i connected to node j;
w_{ij} is a weight associated to the link connecting node i to node j.

In the SiMA-C model the activation value set by trigger activation is used as an initial output value for all associated data. The input value, then, is dependent on (1) the received activation value, (2) the previously received activation, and (3) the association strength. To set the activation value Formula 5.1 is used, where a is given by the received activation value, weighted by the association strength between the two data nodes. The stopping criteria for the spreading process is given by a spreading-deepness-value, which is based on the agent's current unpleasure and reduced after every spreading step (WSG+13).

Two kind of spreading processes are distinguished. Data structures that are associated by explicit associations (e.g., between thing presentations or between goals) have a stored association strength and require more spreading-deepness-value than spreading in implicit associated data, which are treated as having an association strength of 1. This separation resembles different types of association and favors spreading in data that is associated by definition (e.g. a goal must be associated to a thing presentations and action) from optionally associated data (e.g. between norms and goals).

In case of spreading activation in norms, activation is constrained to descriptive norms (identified by their 'type', see Figure 5.4). Hence, two kinds of norms are distinguished based on the conceptual separation of injunctive and descriptive norms (e.g. (CG04)). These corresponds to the psychoanalytic SiMA concepts of internalized super-ego rules in the primary process (injunctive norms) and social rules in the secondary process (descriptive norms). However, the separation is only implicitly considered in two different activation possibilities of norms. Injunctive norms are activated by the current situation through triggered or spread activation - as any other goal. Descriptive norms are activated if their execution is perceived in the external world. In particular, activation happens in the process of salience activation (see above). For instance, in the 'Social Prompintg' exemplary case (see Chapter 6) perceived norms get their salience (and hence activation of the according memorized descriptive norm) based on (1) the memorized valuation of the message-sender, (2) the relationship (i.e., association-strength and type) to the thing presentation 'SELF', and (3) the norm's appreciation by others (represented by message's number of 'social-media-Likes').

Overall, due to the activation process all possibly relevant data for the current situation are processable. Every activated goal is subsequently considered as a possible and appropriate goal

for the current situation. Overall, the activation process impacts decision making only via the next process that it triggers: goal valuation.

5.4 Valuation

The essence of memories is to form expectations about the purpose and effects of goals, i.e., thing presentations and actions, providing information about their value for the agent. These expectations are updated with the goal's feedback on the agent's bodily and normative demands.

The essence of valuation is to consider these memorized feedbacks and adapt the resulting expectations on the demand's current relevance and the goal's applicability. Every goal has memorized valuations representing information about the expected purpose for the agent. Hence, a memorized valuation is the assignment of an unpleasure and a pleasure value to a goal that (1) harmed the body or satisfied a bodily demand (bodily valuation), (2) fulfilled a commandment or broke a prohibition (normative valuation).

These valuations are short-termed, i.e., they value the direct feedback associated with actions and thing presentations. In addition to that, every thing presentations has a memorized summary valuation. It provides an integrative holistic information about all aspects of how an thing presentations changed the agent's state, directly or indirectly. The summary valuation subsume all (goal) experience the agent had with it and considers the whole picture' in providing a summary valuation: it integrates the impact on bodily and normative demands, and their conflicts; long-term feedback, the context. Besides providing information about a thing presentation's purpose, it can be seen as an indication for general approach or avoidance.

Valuation is the process of determining the expected relevance of goals for the agent's current bodily, normative, and perceived situation. If possible goals are activated, a new valuation process is triggered that weight the memorized valuation with the current activation value to determine the current relevance of the goal. That is, only goals are valuated, since valuation is for doing. However, since goals consist of and are activated - amongst others - by associated thing presentations, their relevance is also considered.

Valuation is done depending on the corresponding activation source type (percept, norm, drive, emotion). Goals have memorized normative and bodily valuations and are associated to thing presentations, which have memorized summary valuations. If a goal is activated, memorized valuations are weighted with the goal's activation value from the corresponding activation source type. In case of activation from drives, the goal's memorized bodily valuation is weighted with the corresponding activation values. If a percept or emotion activates a thing presentation, the associated summary valuation is weighted. In case of activation from norms, the goal's memorized normative valuation is weighted. But since norms are themselves only activatable by other valuation sources a normative valuation always implies at least one other type of valuation (bodily or summary valuation, see Figure 5.6).

Hence, the impact of memorized valuations is dependent on the activation value, which represents the relevance of data for the current situation. Therefore in the process of spreading of activation information about the valuation sources are considered.

After the valuation is conducted, a goal's *valuated relevance* is set (or extended in case of previous valuations). Following the principles of the primary process, unpleasure and pleasure (i.e. the pros and cons) are not weighted against each other, but only the dominant one is chosen to set

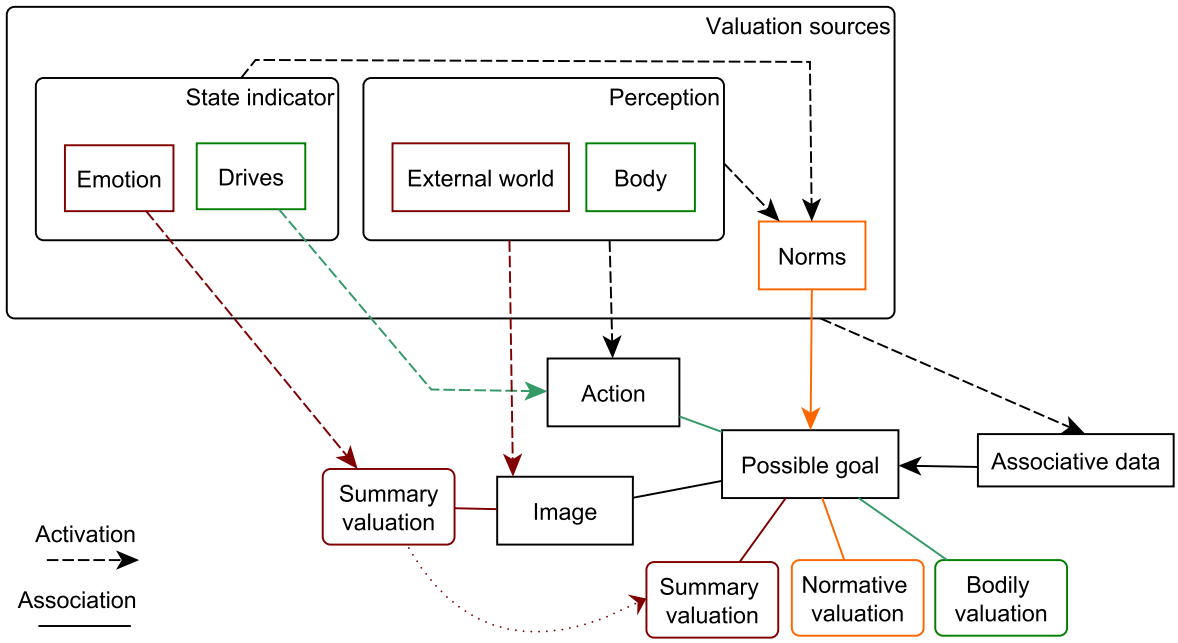

Figure 5.5: Activation and valuation in the SiMA-C model.

the relevance value. No-dominance of unpleasure or pleasure would increase the goal's ambiguity, which increases the necessity for evaluating the goal, where both factors (amongst others) are considered (see Chapter 5.6).

5.5 Conflict Mediation

(Simulated) humans are conflicted beings (see Chapter 4.2.1). The human mind is not able to satisfy the various conditions that the different (and often contradicting) demands and affordances set. Conflicts may occur between different goals, but also within a goal in case of contradicting valuations (e.g., dominating unpleasure in bodily valuation of a goal, but pleasure in its normative valuation). All conflicts occur dynamically in the process of the various inter-playing decision factors. Hence, opposed to valuations, conflict values are not memorized (except in summary valuation).

The main source of conflicts is normative. All activated norms demand their fulfillment. If a norm is activated and not fulfilled by a goal, i.e., no association between the norm and the goal exist, the goal is marked with a norm conflict.

Another source of conflict is a discrepancy between wish and reality. If a goal is physically executable (i.e., not a strategic goal) and highly valuated due to current bodily and/or normative demands, but the affordances of the current situation does not enable its execution, the goal is marked with a reality conflict. Associated goals that are executable in the current situation are activated, as a result of conflict mediation (possibly already in spreading of activation).

The reality conflict type may be extended and specialized by scenario-specific conflicts. For instance, in case of an ownership conflict: If a highly valuated goal includes an object that is not

reachable since it is perceived to belong to someone else, or the highly valuated goal consists of an perceived action that is executed by some one else (variable 'action subject' in Figure 5.4), the goal is marked with an ownership conflict.

Conflict occurrence triggers mediation. Additionally, conflicts are considered as weighting factors in the evaluation process (see Chapter 5.6). Different mediation mechanisms are possible, dependent on the agent's configuration (i.e., personality), conflict type and intensity, and the agent's current ego-strength (represented by the neutralized intensity).

An example for a widely used mediation mechanism is *sublimation*. If an agent has learned that two goals are able to satisfy demands similarly, they are associated in memory. If a goal is conflicted due to the current situation, a mediation by sublimation will go through associated goals and check their conflict potential in the current situation. The goal with the highest association strength, lowest conflict intensity, current activation and valuation values is then chosen to compensate the original goal. This includes displacement of valuations from the conflicted goal to the sublimed goal, dependent on their association strength and current ego-strength. Another example of a mediation mechanism and a possible extension of sublimation that is used in the exemplary case 'Social Prompting' is imitative identification. For instance, if no appropriate sublimation in memory is possible and a perceived goal of another agent could be appropriate, this mediation mechanism would lead to imitate that behavior by changing the 'action subject' variable (from the perceived agent to the thing presentation 'SELF') depending on the valuation of and relationship to that agent.

5.6 Evaluation

The SiMA-C model is a dual processing model of the human mind. The two processes can be regarded as unconscious and conscious or as primary or secondary process. However the second is dependent on the primary process and is only a extension of it. Both follow different principles, but operate on the same values. The primary process activates and valuates data and tries to mediate in case of conflicts. However these processes only consider local conditions, e.g., valuations according to a specific demand. If necessary and possible, the second process uses a global view in considering all factors. After the primary process' memory-based operating principle, the secondary process uses reflective operations in explicitly relating the agent's current state to the expected state represented by the goal's valuation.

Overall goal selection in the SiMA-C agent is based on a goal's relevance value. The determination of this value is a stepwise process. In the primary process valuation and its manipulation sets the relevance value, which may be extended in the secondary process by evaluation. The used terminology reflect the functionality of the processes of valuation and evaluation, and also emphasizes their relation: evaluation includes and extends valuation. This corresponds to the assumptions that the secondary process cannot change the primary process' goal's valuation, but only uses it differently (other operating principles). This is immanent to the concept of the unconscious in the primary process.

According to the principles of the primary process, the different valuations are just aggregated to a *valuated relevance* (using non-proportional aggregation, similar to 5.1). In the secondary process *evaluated relevance* is calculated in weighted mult-criteria aggregation. Evaluation is able to use all information gained sofar in the primary process to decide the most relevant goal in the current situation, in particular to decide which goal is overall best for enhancing an agent's

state. However, in considering bounded rationality, evaluation follows an approach of satisficer, not optimizer. In this regard evaluation is done gradually depending on two dynamic factors (see Figure): (1) the necessity for evaluation represented by the clarity of goals' relevance (i.e., small distance between goals with the highest relevance), and (2) the possibility represented by the agent's neutralized intensity.

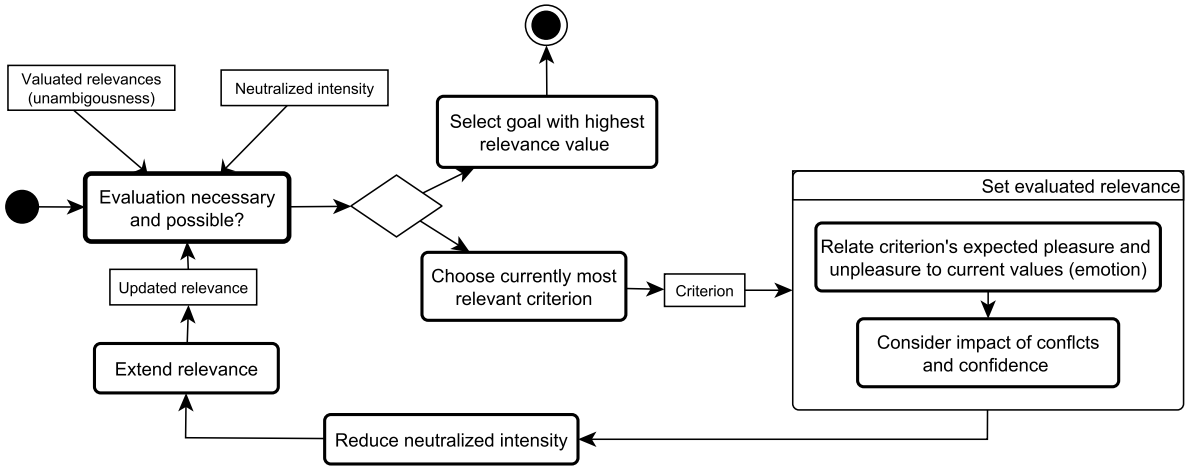

Figure 5.6: Evaluation in SiMA-C (UML activation diagram)

6 Simulation

What I cannot create, I do not understand

Richard Feynman, 1918-1988

After the conceptual and functional model is developed with the support of the exemplary cases, the application of the model in the cases' area (environmentally friendly behavior and social interactions) is tested in different scenarios. This entails validating the model by comparing it against expectations from the underlying psychoanalytic and neuroscientific assumptions, and against empirical data. The former validates the consistence and explanatory abilities of the model, the latter additionally validates its prediction abilities in the application domains. In both cases the plausibility that the underlying theories represent the functionality of the human mind is corroborated.

As described in section 2.3, the central test specification is given by the simulation cases. Evaluation consists of comparing the simulation against the expected resulting behavior of agents and against the expectations of how this behavior is generated and determined. This includes testing the model's expected determinability, in particular if a change of behavior determinants results in the expected change of behavior. As described in Chapter 2.3, if the expectations from the simulation case specification are not met in simulations, an analysis including an interdisciplinary review is done. This resulte in theory, model, implementation, or parameter development, and in worst case falsification (see Figure 2.2 and 2.4).

As mentioned, the thesis' question is tackled in two different domains (environmentally friendly behavior and social interactions), in two different simulation environments (Artificial Life and Decision Simulation Environment). Therefore, the simulation case specifications and results are described in separated chapters.

6.1 Artificial Life Simulations of Social Interactions

The exemplary case 'Two agents and a food source' (see Chapter 4.1) is used to tackle the question of social interactions, in particular why agents act prosocially or aggressively. After specifying the exemplary case's and empirical data's expectations into simulation cases, i.e., specifying which parameters (determinants) are expected to generate a specific action, the expectations are tested in simulation.

6.1.1 Development and Evaluation with Incremental Simulation Cases

Usually an exemplary case (EC) includes the demonstration of multiple concepts (see Chapter 4.1). In the context of an holistic approach, this increases the complexity in developing and evaluating the model at hand, due to the interaction of a high number of parameters and variables. In the process of development and evaluation, control of these variables and parameters is necessary[1]. To break down the complexity in development and evaluation, the EC is split into multiple simulation cases (SC) as shown below. This enables incremental and iterative requirements analysis, development, and evaluation: We start with the basic version of the simulation case, and extending it in every iteration with a focused functionality. Thereby we extend and focus on the respective determinants for this functionality. In such a hierarchical SC separation, the SCs on the one hand build upon each other and on the other hand are used to develop and test a dedicated functionality.

Hence, the single SCs corresponds to single simulatable model versions (cf. Figure 2.4) and serve as milestones. In this regard, all SCs (but the last one) are intermediate milestones and only helping constructs (see Figure 6.1). The last SC version includes all model functionalities and determinants and can be used for model exploration, without any prior specifications and expectations.

The baseline scenario is extended with every incremental SC with the respective determinants that represents the new functionality and the corresponding process description. Every SC is concerned with establishing the general and simplest working of the model, e.g simulating the action to eat. After that, the focus-functionality of the SC is then demonstrated in the alternative scenarios (see Figure 6.1), using the new determinants[2].

6.1.2 Simulation Case Specification

The specification of narrative exemplary cases (see Chapter 4.1) into structured simulation cases goes in line with (and supports) the specifications of concepts and development of the model. As described in Chapter 2.3 a simulation case focuses on (1) the specification of determining parametrization ('behavior determinants'), (2) predicting variables ('subjective predictors'), and (3) the specification of alternative scenarios. In particular, this concerns how a change in the baseline scenario's behavior determinants causes a change in the agent's behavior. The simulation case represents all outcomes of the exemplary case that are relevant for tackling the research question and demonstrating the underlying concepts. Another aspect is to demonstrate all relevant mental functionalities in single alternative scenarios. Hence, the alternative scenario and the corresponding determinants is selected in this regard.

These scenarios are used for model evaluation in simulation by providing a parametrization scheme and expectations on the predictors and decided action. Additionally, model specification - enabled by simulation cases - supports the increase of unification in a functional model. Another benefit of simulation cases in terms of calibrating a function model is enabling model parameterization in a well-structured manner.

[1] Of course, after we have validated that the simulation model behaves as specified, we dismiss control of parameters and variables to obtain model insights.

[2] Not all simulated actions can be used as good examples to introduce new functionalities, i.e., to show how a new functionality determines an alternative action. However, every SC has to simulate all actions, to enable incremental development of the whole model.

Motivations from Nature and Culture: Eat, Share, Fight, or Flight (Exemplary Case 'Two agents and a food source')
The first exemplary case (EC), as described in Chapter 4.1, merges two exemplary cases due to their common theme. EC 1 ('Adam seeks Schnitzel') aims to demonstrate principal mechanisms for developing the basic functions of an artificial mind. EC 2 ('Adam and Bodo') further specifies EC 1 and extends it by including the topic of emotions, their bodily expressions and interpretation by the other agents. Given the commonalities of the two exemplary cases, in the following, the structuration into simulation cases is focused on EC 2.

SC separation and overview

The overarching topic in EC 2 is to introduce emotion into the SiMA model. To break down the complexity in incremental development and evaluation (see above) EC 2 is separated in three coherent and hierarchical simulation cases (see Figure 6.1). The separation follows the demonstrated functionalities of emotion.

SC 2.1 focuses on the different (e.g., reactive and reflective) valuative and evaluative usage of emotion in decision making. It introduces emotion for holistic and integrative valuation and demonstrates its interplay with other valuation mechanisms in determining the agent's behavior. This includes fine-adjustment of valuations from drives, activating and valuating memories as a summary valuation. It also considers actions and plans that are not directly activated and valuated from drives, by considering memories activated from perception, and unpleasure avoidance (demonstrated by a new alternative action - leaving). Additionally the SC shows how emotion operationalizes social norms via their valuation functionality.

SC 2.2 is concerned with the bodily expression of emotion and its impact on other agents. It shows the functionality to represent the internal state externally, and adapts it on others' expressed internal state. Recognizing the bodily state of others as emotion not only enables the recognition of subjects, but also serves (unconscious) bonding and implicit behavior adaptation on other agents and their current state. Overall, the SC shows how bodily expressions serve an unconscious coupling and adaption between agents. This can be termed as affective empathy.

SC 2.3 focuses on using emotions to interpret others' intentions consciously (i.e., in the secondary process). Additionally, action recognition is considered in decision making. Hence, in SC 2.3 emotion is used as feelings under consideration of a conscious adaption on an other dynamic agent, often termed as mindreading. This simulation case is the first SC with two fully active agents. Due to increasing the complexity and controllability (of the external environment) Bodo was passive in SC 2.1 and only displaying bodily expressions in SC 2.2. Uncontrolled external environmental and dynamic agents demonstrate the combinatorial possibilities of the various determinants and their impact. SC 2.3 is the final simulation case in the hierarchy and hence includes all functionalities of SC 2.1 and 2.2. Due to hierarchical model development and provision of basic functionalities, after simulating the previous simulation cases, the dynamic and complexity of two active agents can be tackled in SC 2.3.

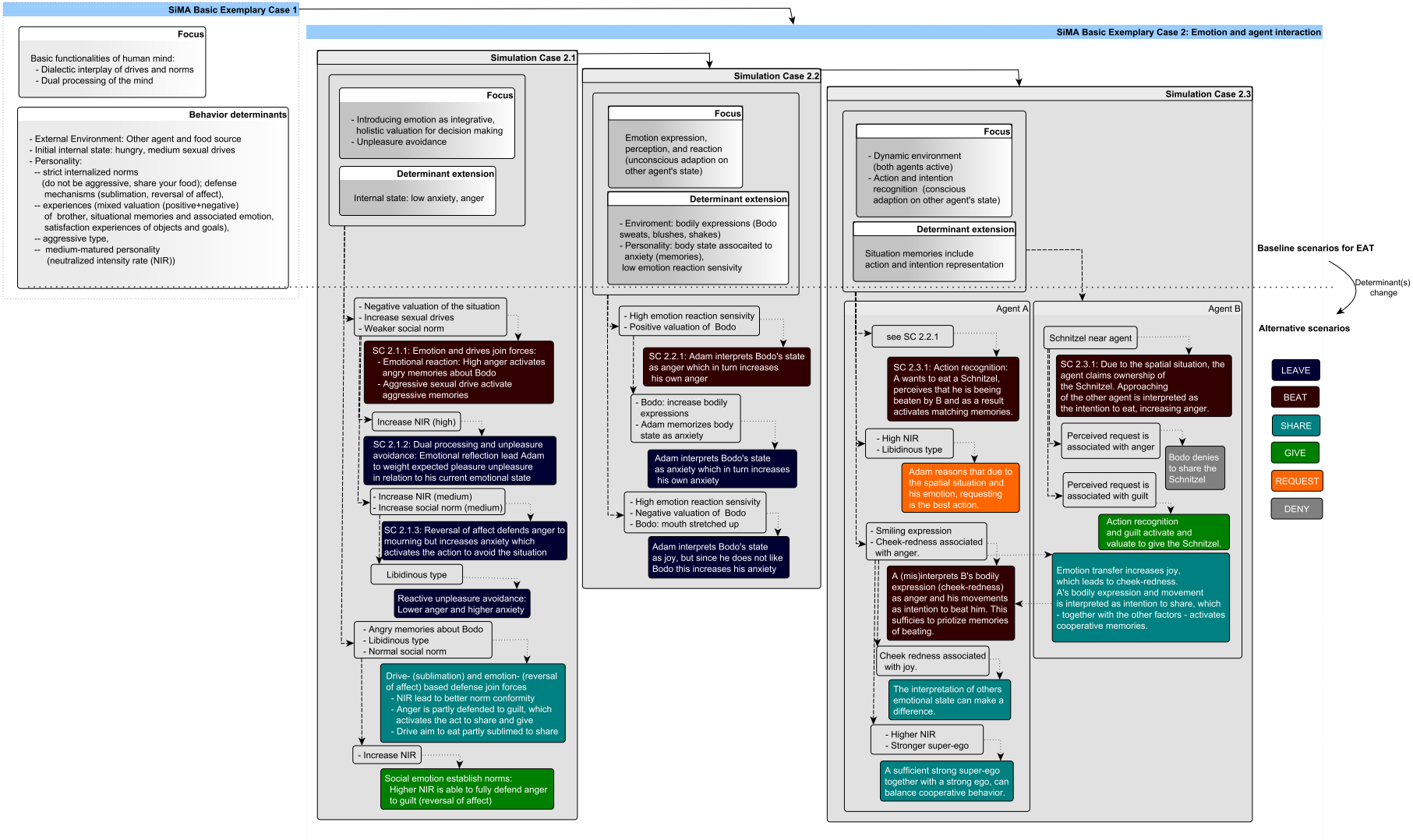

Figure 6.1: Incremental simulation cases (SC) specification for development and evaluation of EC 'Two agents and a food source'. Four groups of parameters (behavior determinants) are expected to determine the agent's behavior: The initial internal state (bodily needs, preactivated memories), the external environment, memories (valuated experiences and norms) and other personality parameters. Alternative scenarios specify how a change of these determinants (transparent boxes) is expected to change the agent's behavior and specify the reasons (colored boxes). This SC structure serves a guideline for incremental model development and a test template for evaluation.

6.1.3 Simulation Results

The hierarchical simulation cases (see Figure 6.1) are used for an iterative development and calibration. Hence, the simulation is tested against the expectations specified together with psychoanalysts. In case of a deviation, software testing is done first. After that, the parametrization and model specification is revised in an interdisciplinary review (see Figure 2.2). The successful simulation of all described scenarios validates the knowledge translation into a computational model and corroborates the model assumptions.

Next, an overview of the scenarios is given by sketching the different importance of decision factors. After that, we focus on demonstrating the different reasons for the agents' action selection, in particular the factors that may change action selection. This leads us to answer the question about the determinants of cooperative behavior in the SiMA agent. To give an insight in the internal workings of the agent's decision making process, visualizations (so called inspectors) are used. They help to comprehend the causal chain in decision making, i.e., to demonstrate how determinants cause behavior through a sequence of functions.

6.1.3.1 Basic actions are directly determined by drives and memories: Eat

The baseline scenario examines the impact factors for the probably most fundamental body-related action in life: eating. With drives as the focus in this scenario, it provides the basis for all subsequent scenarios, since it considers the basic mechanisms for motivation and autonomy in humans. The basis-functions for this scenario were already provided by (Deu11) and further specified in the thesis at hand (see Chapter 4.3.2).

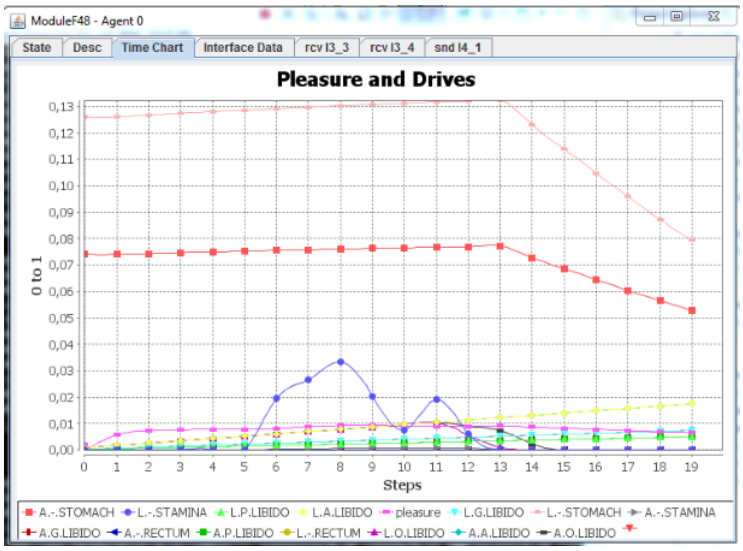

Figure 6.2: Drive evolvement in the eat scenario 2.2.0. (Screenshot, Scale of simulation steps 1:10) Competition of drives and the satisfaction of libidinous hunger (in rose) and aggressive hunger (in light red). Searching for food increase the stamina drive (in light blue), but after finding the Schnitzel the drive decreases. After eating both drives are reduced to higher stomach fill level.

The main determinant of eating - obviously - is hunger, sourced from the agent's organ state and represented by a drive (see Chapter 4.3.2). Some personality parameters determine how

the drive's quota of affect (QoA) evolves in time. However, besides the value of the quota of affect (see Figure 6.2), the main determinants of eating are given by an agent's memories about hunger satisfaction and the opportunity to use the memories in the current environment. Hence, following mentioned drive-based factors, which are in a dependence-relation with the drive's QoA, determine their priority: matching between QoA and memorized drive satisfaction, availability of drive object, effort of drive aim, and the possibility of multiple drive satisfaction. Additionally to the drive-based determinants of eating, emotion could enforce any goal, e.g. eat could be associated to joy. However, given psychoanalytic specifications, in our parametrization of the determinants, emotion only play a secondary role in eating.

Figure 6.3: Simulation of the SC scenario 2.2.0 - eat. The emotional state reflect the drive state and the memorized emotion associated with the current external situation. Eating the Schnitzel (Step 15) reduces the hunger drive, the resulting unpleasure reduction also reduces anxiety, anger, and mourning. The bottom right figure shows the competition between all possible goals in the current situation. All decision factors considered, the eat goal wins slightly.

The high valuation of the goal to eat is only reduced minimally by the defense mechanism sublimation (moving the valuation to the share goal) due to insufficient neutralized intensity. However, the final selection of the eat goal is also due to the low valuation of other goals. If, for instance, another drive would be high enough, the reduction of the valuation of eating by the defense mechanism would suffice. This emphasizes that not the absolute value of a goal's valuation is

crucial, but its relation to other goals.

After eating the Schnitzel the agent (simulation step 12 in Fig. 6.2) not only the satisfied drives decrease accordingly, but also the agent's emotional state changes due to the changes of unpleasure and pleasure (see Fig. 6.3).

The concrete values of the relevant determinants are highlighted in Fig. 6.4.

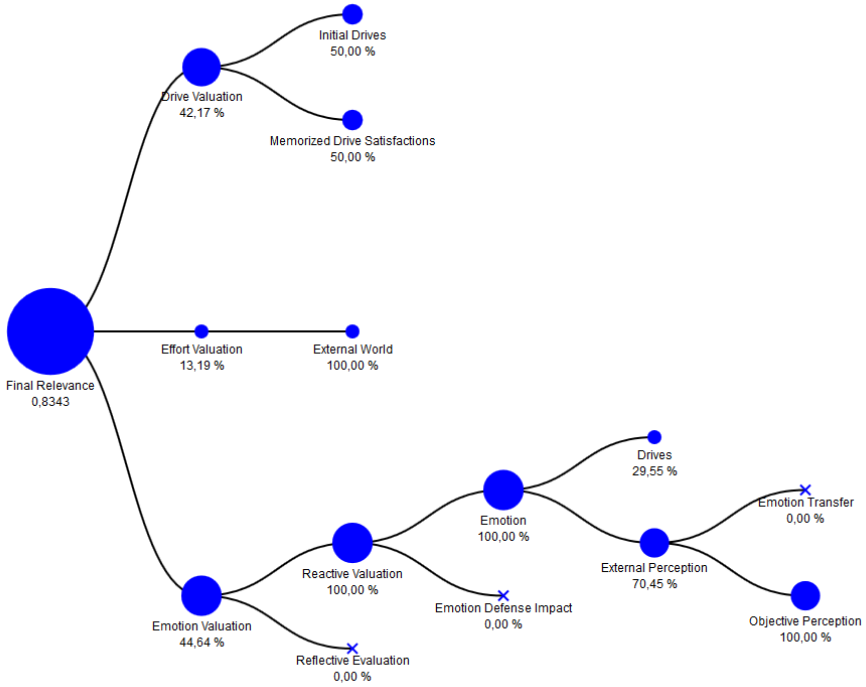

Figure 6.4: Basic behavioral determinants, exemplified with the eat SC scenario 2.2.0: The final relevance of a goal for the current internal and external situation is based on its valuation from drives and emotion. Drives are the most important impact factor in the eat scenario, directly due to drive valuation (42,17%) and indirectly via emotion (29,55%). Due to low neutralized intensity, emotion are used reactively in valuation, i.e. by using memorized emotion that are activated by current perception. Due to the memorized emotion activated by perception, drives have the lower impact on the current emotional state (29,55%). Since emotion is not defended by super-ego rules (social norms), valuation only uses basic emotion.

After looking at the baseline action eat, next, we analyze the other reactive action of the simulation cases: beat. After that we take a look at factors that enforce deviations from the basic reactive actions of eating and beating, i.e., we analyze the determinants of cooperative behavior.

6.1.3.2 An interplay of drives and emotion is required: Beat

The determinants of the beat action can be separated into drive-based and emotion-based determinants. Hence, the beat action is the first (reactive) action, where emotion plays a significant role. In the end - putting extreme cases beside - only the dependent interplay of drives and emotion leads the agent to beat the other one (see Figure 6.5). Hence, only by the aggregation of different activation sources - drives and emotion - with different activation patterns, a sufficient activation and valuation of the beat goal is reached (compared to other currently possible goals

for the agent). Since simulation case 2.2 incorporates 2.1 (see Figure 6.1), we can use it to analyze the beat action.

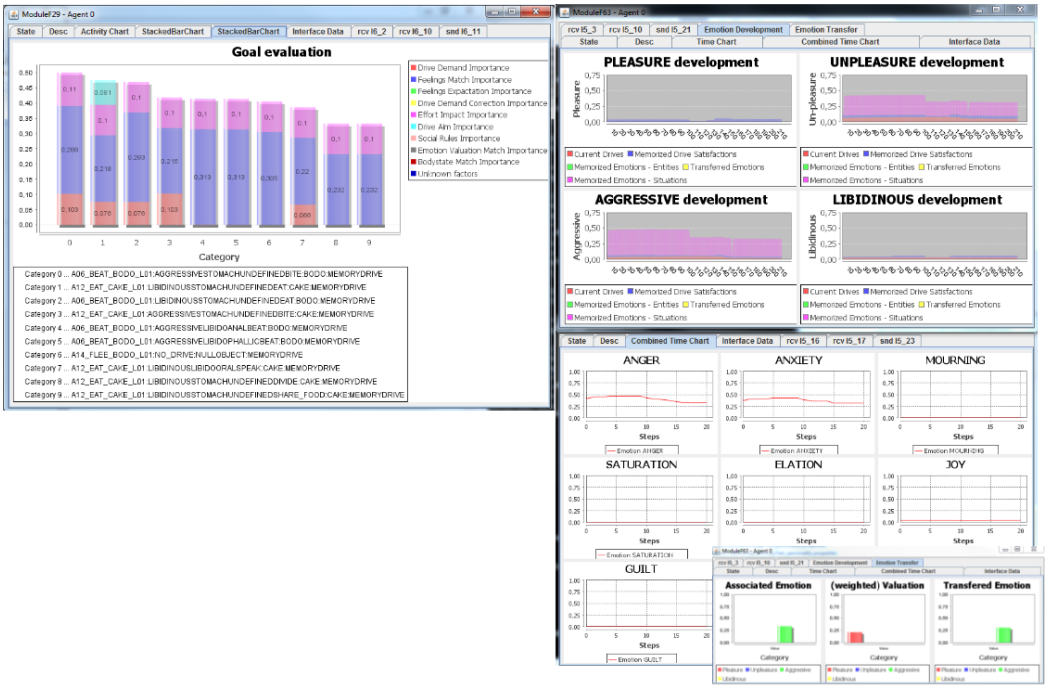

Figure 6.5: Simulation of beat scenario 2.2.1 (screenshot). The other agent's body expression lead to beating. Increasing emotion reaction sensitivity and positive (!) valuation of the other agent suffice to get the green agent to beat the red agent instead of eating the Schnitzel. The mid right figure (yellow stripe) shows that the main source of the aggressive emotion component comes from emotion transfer. The bottom right figure shows that the red agent's bodily expression is associated with aggresive unpleasure (i.e., anger), and since the agent is valuated positively, anger is transferred to the green agent's current emotion. The left figure shows how only the impact from emotion together with drives lead the beat goal (most left stack) to outcompete the eat goal. Changing neutralized intensity would lead the agent to leave the situation instead of beating the other agent (see Fig. 6.13).

Drive-based determinants of the beat action are given by sexual drives. Following psychoanalytic theory (see Chapter 4.3.2), sexual drives are not confined to genital aspects. In this regard, the aggressive phallic sexual drive - dependent on the personality, i.e., memories about how to fulfill aggressive sexual drives - is satisfied by attacking other agents. But the simulation shows that even with high sexual drives an agent would not choose to beat the other agent, since the hunger drives compete with sexual drives and defense mechanisms defend against the forbidden goal to beat - even with a medium ego strength (dependent on neutralized intensity rate and drives). Also, the eat goal satisfies both, self-preservation and sexual drives (multiple drive satisfaction, see Chapter 4.3.2) and hence would be preferred in this regard. Only if (1) the parameter to strengthen the super-ego rule (i.e. social norm) not to beat other agents is decreased, (2) sexual drives are increased, and (3) the memorized unpleasureful valuation of similar situations (i.e., fighting with the brother about a food source) is increased, the determinants lead the agent to beat the other one (cf. SC 2.2.1 in Figure 6.1). The concrete values of the relevant determinants are highlighted in Fig. 6.6.

With the additional consideration of emotion-based determinants, the beat goal is also selected in non-extreme situations (i.e., extreme high sexual drives). However, since drives are a main

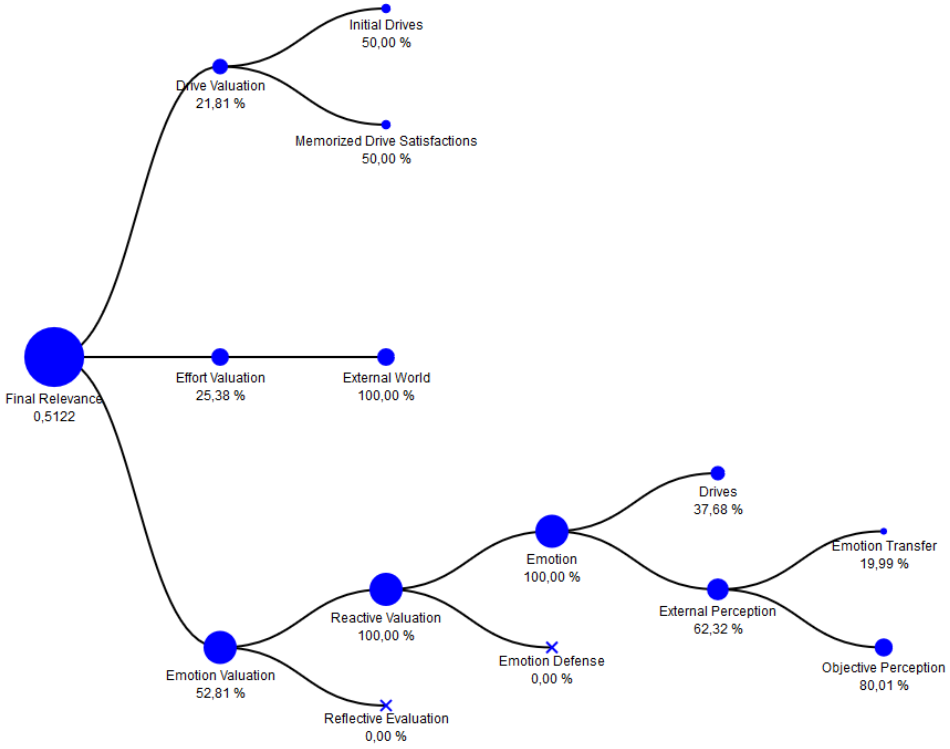

Figure 6.6: Determinant values of the BEAT scenario 2.2.1. Over 52,81% of the beat goal's relevance value is determined by emotion, from which 100% is due to reactive memory activation by the agent's current perception and emotion, which are mainly determined (62,32%) by activated memorized emotion by perception (80,01%) and also by emotion transfer (19,99%), i.e., attributing emotions to the other agent's bodily expression and getting influenced by that.

source for emotion, these two decision factors cannot be considered separately. But in the case of the beat goal, the main source for the agent's emotion are memories triggered by perception (see Figure 6.5, cf. Figure 6.1). Perceiving the other agent and a food source activates memorized images about fighting with the brother (who looks similar to the other agent), which are valuated with anger. However, the impact of sources differ in the four emotion factors (see Figure 6.5).

The central role of activated memories for the agent current emotion cannot be changed by increasing the personality parameter of emotion reaction sensitivity and - keeping the unpleasureful memories about fighting with the brother - by increasing the memorized pleasureful valuation of the other agent (see Figure 6.5 and cf. scenario 2.2.1). Usually a perceived situation activates multiple memorized images, and hence *multiple* memorized valuations affect the agent's current emotional state. The *single* attributed and transferred emotion (cf. Chapter 4.3.3.8) is only able to provide a minor source for the agent's current emotion - even if the agent is very sensitive to other agents' emotional expression. However, as shown in scenario 2.2.1 - in the right constellation (and competition to other actions) - it suffices to get the agent to beat instead of eating the Schnitzel, since emotion transfer only increases the aggressive emotion component but not the other emotion components, and hence anger rises (see Fig. 6.5).

Another determinant of the agents' behavior is the spatial configuration of the external envi-

ronment, introduced in SC 2.3.1 (also see 6.11). In particular, the Schnitzel's proximity can be interpreted as ownership (memorized images are extended in this regard). In this regard ownership is used to extend the activation pattern of similar memorized images. That is, the perception of (spatial) ownership requires a respective association type (ownership) for a perfect matching between perceived and memorized image.

SC 2.3 - as the first simulation case with two active agents - also introduces dynamic determinants, representing the dynamic aspect of the other agent's behavior, in particular action and intention recognition.

In SC 2.3.4 - given the baseline configuration of the external environment - agent A, although initially in a joyful mood, misinterprets the other agent's emotional state as anger and his intention as aiming to beat him, which increases his anger (see Fig 6.8). The resulting anger - together with other factors (the interpreted intention of the other agent, and - although only to a small degree - aggressive sexual drives) - activate memorized images of beating. At the same time agent B aims to share, co-determined by an emotion transfer of agent A's joyful expression and cooperative intention interpretation. For the simulation sequence and the inner workings see Figures 6.7 and 6.8.

Figure 6.7: Simulation sequence of the beat scenario 2.3.4 (screenshots), with two active agents. Read from top left to bottom right. The joyful red agent approaches the Schnitzel (picture 2). Due to its determinants the red agent decides to share the Schnitzel with the green agent (pictures 3 and 4). But the green agent miss-interprets the other agent's bodily expressed joy as anger and the movement towards him as intention to beat him, which increases his anger and activates the goal to beat the red agent (picture 6). To see how the determinants lead to this behavior see Fig. 6.8

Only the correct recognition that the other agent got angry and will beat him would activate memories of beating (back) (cf. SC 2.3.1 and see Figures 6.10 and 6.11). In this regard one can speak of a self-fulfilling prophecy. These two scenarios (2.3.4 and 2.3.1) also emphasize that external determinants (such as body expressions) only have an impact via internal determinants (such as the interpretation of the body expression).

The same determinants (action and intention recognition) are able to determine cooperative behavior, given other contextual determinants (see Chapter 6.1.3.4).

Figure 6.8: Simulation of the beat scenario 2.3.4. with two active agents, where a miss-interpretation leads to beat the red agent (Adam), although Adam shares the Schnitzel. The top figures show the emotional state of the red agent, and the bottom figures those of the green agent. The red agent is joyful, mainly due the pleasureful memories about his older brother (activated by perceiving the similar green agent) and due to emotion transfer caused by the interpreted joyful state of the green agent (see yellow stripe in pleasure development, top right figure). Sharing the Schnitzel (simulation step 90) increases the joy, but getting beaten also increases anger and anxiety (step 10). The top left figure also shows the pleasureful emotion the red agent associates with the green agent's body expression, but since the positive valuation of him is low, the emotion transfer is not proportional to the attributed pleasure. For the red agent's personality in this scenario sharing is determined by activation of cooperative memories by joy and the (miss)interpretation of the green agent's intention to share (cf. the share scenario determined by defense mechanisms and guilt).

Overall, the beat scenario demonstrates the different sources and mechanisms of emotion generation: drives, memory activation, and emotion transfer (i.e. affective empathy) and their interplay. Thus, it shows how emotion serves as an integrative holistic summary valuation.

6.1.3.3 Unpleasure avoidance and reflective behavior: Leave

Emotion does not only operate in a reactive manner, but also facilitate reflective thinking. In the secondary process of the SiMA model they also may operate in a reflective manner (see Chapter 4.3.3.5) *as* feelings (see Chapter 4.3.3.4). This is dependent on how mature the personality is (represented by ego-strength, i.e., the personality parameter 'neutralized intensity ratio' and drives) and the current drives ('neutralized intensity' is a share of drives, see Chapter 4.3.2.4). This functionality enables the agent to reflect about the current situation, i.e., using emotion

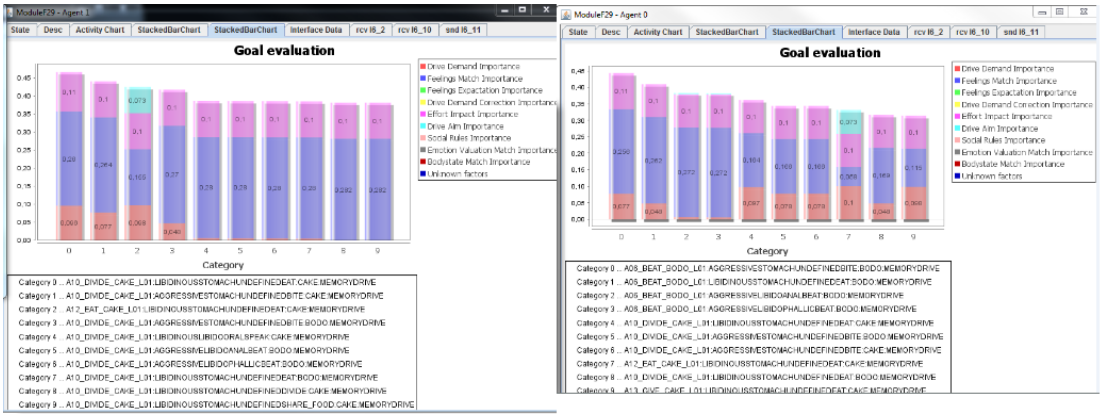

Figure 6.9: Simulation of the beat scenario 2.3.4. For both agents the decision is quite distinct: The red agent's first 10 prioirized goals - except for one - (left figure) are about sharing (dividing is the first goal of sharing), only afterward followed by eating. Even if the green agent's prioritization are also clear, he has 'mixed feelings': the goals to beat the other agent are mixed with goals to eat and to leave the situation (right figure).

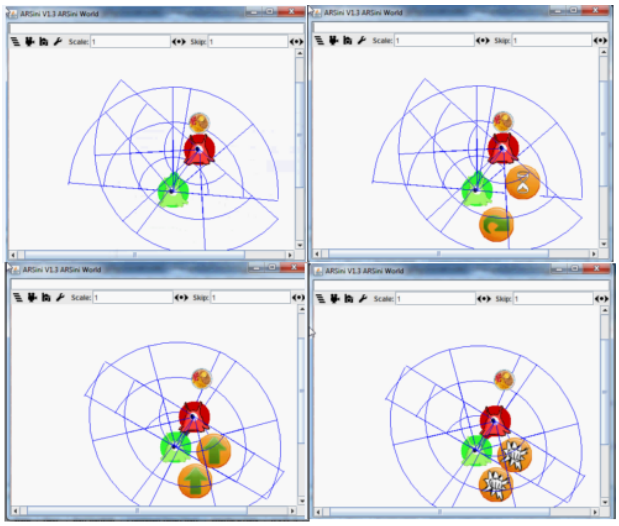

Figure 6.10: Simulation sequence of the beat scenario 2.3.1, with two active agents and correct interpretations of the others' emotional expression and intention. Read from top-left to bottom-right. The green agent wants the Schnitzel and approaches it. His movement towards the red agent's Schnitzel together with the green agent's angry expression (due to high drive-based unpleasure) lead the red agent to approach him as well for the sake of beating (the beat goal is activated due to increasing anger and the interpreted intention of the green agent). Consider that the spatial situation of the Schnitzel clarifies ownership and adds to the activation of according images and hence goals.

as feelings: What are the consequences and how does the negative consequences relate to the positive ones? And most importantly: How to relate these consequences to the current state (i.e., the current emotion)? Hence, this scenario emphasizes on the different valuation mechanism of the agent's internal state: activation of memories by the agent's current internal state (emotion) and reflecting on the internal state (feeling).

As seen in the specification of the SC scenario 2.1.2 (see Figure 6.1), starting from the determinant setting of the beat scenario, a change of the personality parameter 'neutralized intensity ratio'

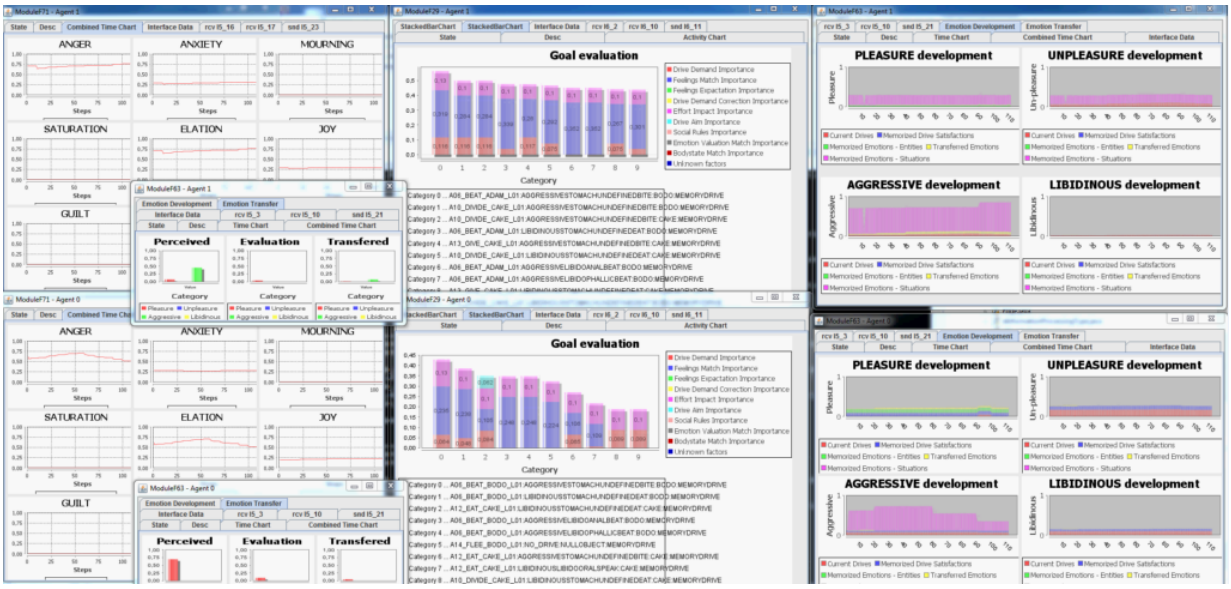

Figure 6.11: Simulation of the beat scenario 2.3.1. with two active agents and correct interpretations of the others' emotional expression and intention.

Figure 6.12: Simulation sequence of the leave scenario 2.1.4. Read from top-left to bottom-right. Unpleasureful memories activated by perceiving the red agent and emotion transfer due to the body expression lead to increasing anxiety in the green agent and reactive activation of the goal to leave the situation. Note the gradual increasing expression of emotion (e.g. blushing) in the green agent, indicating a rise of anxiety.

suffices to get the agent to leave the situation instead of beating the other agent. Of course - as in the beat scenario - the agent is driven by aggressive drives and is angry due to them. As described above, this is mainly caused by the unpleasureful memories of the other agent, and by

emotion transfer; but due to the high neutralized intensity the agent is able to reflect about the situation and decides an action - to leave - which is beneficial for his current state, i.e. avoids unpleasure and neglecting the minor pleasure expected from the aggressive sexual drives. That is, the agent's state and memorized valuation is the same as in the beat scenario - only how the agent processes the situation and its valuation differs (see Fig. 6.13). Although in this scenario it seems that drives do not play a role, since the goal to leave is only chosen due to the agent's feelings and reflection on it, this is not the case: The main determinant in SC 2.2.1 for the agent to prefer leaving instead of beating is given by the personality parameter neutralized intensity rate (enabling a reflective mode on the agent's feeling), which is dependent on drives' quota of affect to have an impact. Also, drives are a central source for emotion (see Chapter 3.3).

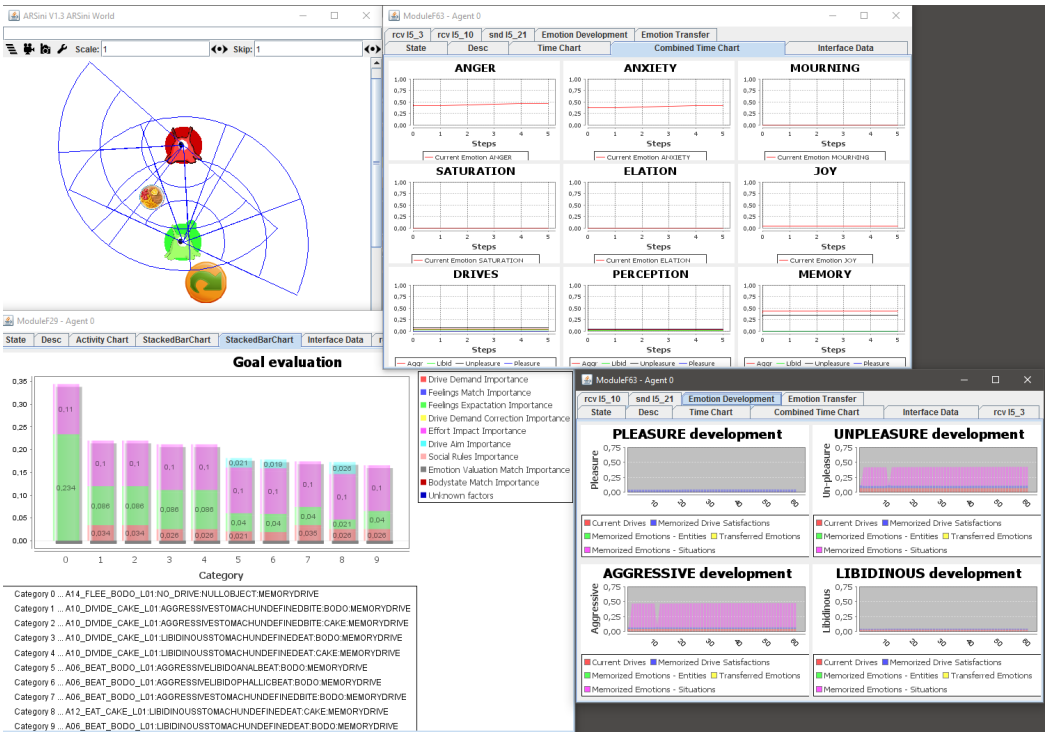

Figure 6.13: Simulation of the 2.1.2 leave scenario demonstrates reflective instead of reactive use of emotion as feelings. Although the emotional state is the same as in the beating scenario 2.2.1, the agent decides to leave, since the higher neutralized intensity allows for using emotion as feeling by relating the agent's current state to the expected state from valuation. The bottom right figure shows the high impact of the reflective mode of feelings (in green, opposed to blue in other scenarios).

In the SiMA model reflecting on the situation is not the only way of choosing to leave instead of beating the other agent. As shown in an alternative scenario in SC 2.1.3 (see Figure 6.1) and in Fig. 6.14, the agent does not have to be ' an intellectual' to prefer leaving instead of beating. Starting again from the beat scenario, increasing the strength of the super-ego rule not to beat other agents (i.e., social norm), together with an only medium high neutralized intensity ratio, suffices to get the agent to leave the situation instead of beating the other agent. This is due to the defense mechanism 'reversal of affect', which moves intensity from the aggressive to the libidinous emotion component. However, since the conflict is not high enough mourning is higher than guilt, co-determining the agent to leave instead of sharing the Schnitzel. (A dominance of libidinous emotion component is regarded as mourning, when it is extended with a high conflict

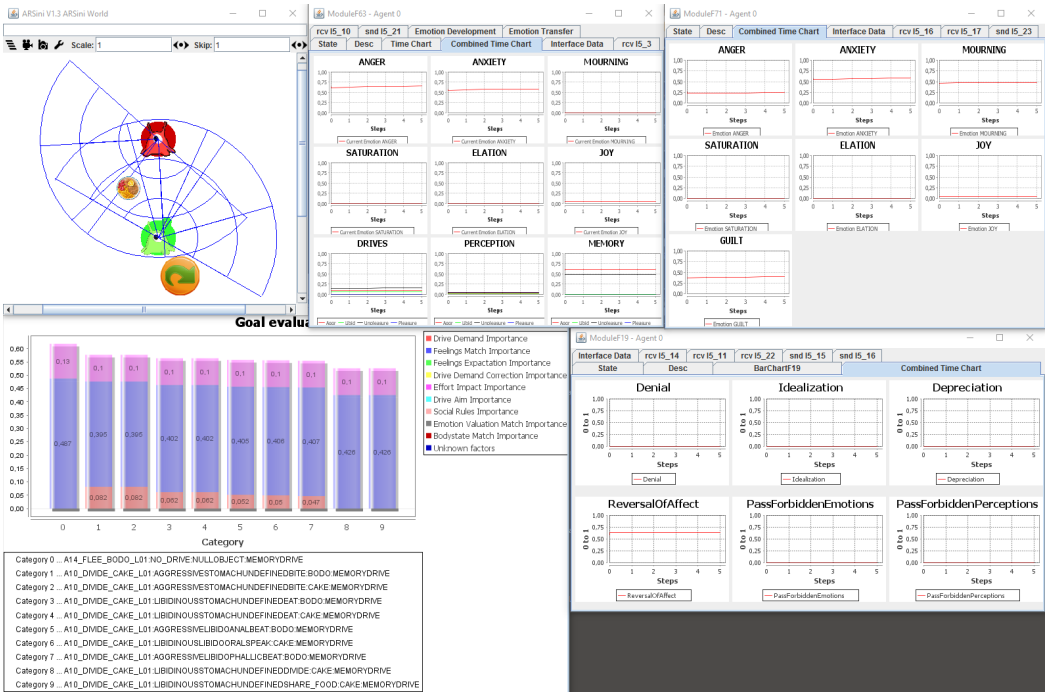

Figure 6.14: Simulation of the 2.1.3 leave scenario. Starting from the 2.1.1. beat scenario, a small change in the neutralized intensity rate suffices to get the agent to leave the situation due to a sufficient working defense mechanism (reversal of affect - resulting in mourning and guilt instead of anger) as shown in the left most figures. The bottom right figure shows that the goal to leave is not valuated by any drive, but by reactive mode of feelings (in blue).

the constellation is regarded as guilt, see Chapter 4.3.3.)

As shown in SC 2.1.4 (see Figure 6.1) and Fig. 6.12 another scenario of leaving can be reached by high anxiety in the agent. Therefore the only determinant that needs to be changed is the personality parameter regarding the separation of drives into aggressive and libidinous drive components (see Chapter 4.3.2). Changing the ratio in favor of a libidinous personality also changes emotion generation (see Chapter 4.3.3): A dominance of unpleasure together with a dominance of libidinous drive components is mapped into anxiety (instead of anger in case of aggressive dominance). The anxious state of the agent activates memorized images to leave the situation.

As before, emotion - in this scenario anxiety - can be increased by attributing an emotional state to the other agent and getting affected by that (see SC scenarios 2.2.1 and 2.2.2). Increased bodily expression of emotion by the other agent leads to emotion contagion in case of positive valuation of the other agent (see SC 2.2 and Figure 6.16). In case of a joyful other agent and unpleasureful valuation of the other agent, opposed emotion contagion takes place, which also increases the agent's anxiety. For a comparison of these two scenarios see Figure 6.17.

6.1.3.4 Defense mechanisms or emotion transfer and intention interpretation are required for any cooperative behavior: Share

Simulations of the exemplary case 1 in the SiMA project (SWJ+14) already show how high ego-strength is able to defend hunger into cooperative actions. That is, in case of high ego-strength

(high neutralized intensity rate in combination with drives) the super-ego rule creates a conflict with the hunger drive, which in turn activates the defense mechanism sublimation, which displaces the drive's valuation from the eat to the share goal.

With the introduction of emotion in the thesis at hand, additional determinants of cooperative behavior are parametrizeable in simulations. This increases the possibilities of determining cooperative behavior in different personalities. Hence, with this extension, simulations of the SiMA model get psychologically more plausible. As depicted in the specification of SC scenario 2.1.5 (see Figure 6.1), drive-based defense mechanisms are supported by emotion-based defense mechanisms. That is, due to the agent's high anger (mainly caused by negative valuations of the other agent, see the beat scenarios in Chapter 6.1.3.2) - i.e., dominant aggression in the emotion vector - the super-ego rule 'do not be aggressive' generates a conflict. Subsequently, the conflict is mediated by the defense mechanism 'reversal of affect' - increasing the libidinous emotion factor out of the aggressive and hence creating guilt out of anger (dominant unpleasure and conflict is mapped to guilt, see Chapter 4.3.3.2). Hence, this scenario is the first one that introduces extended emotion. In the further process, guilt activates matching memories of sharing and giving. Of course, the share-goal is already activated by the drive-based defense mechanism of sublimation, but only with the additional activation from guilt, the share-goal is activated sufficiently to compete with the other actions (eat, beat and give). In this regard the super-ego has an two-fold impact: directly via subliming eat to share, and indirectly via creating guilt, which operates in the secondary process by activating the cooperative actions share and give. Furthermore, one can observe that (for the specified personality in this scenario) indirect causation of cooperative behavior is triggered by emotion, in particular by anger. In this regard, one can say (provocatively) that anger (even towards the to-be-cooperated-with agent) supports cooperative behavior, as long as the agent has a strict super-ego (i.e., the determinants super-ego rule-strength) and enough ego-strength (i.e., the determinant neutralized intensity ratio in combination with drives).

With an even higher ego-strength (i.e., the neutralized intensity ratio) anger is defeated almost completely into guilt, determining the agent to give the Schnitzel to the other agent (see SC scenario 2.1.6). Emotion transfer may increase the anger and hence lead to stronger cooperative behavior in this personality: Higher anger leads to a higher conflict and a stronger reaction by the defense mechanism reversal of affect - turing anger into mourning and guilt.

But as shown in Fig. 6.8 and described in Chapter 6.1.3.2, the goal to share the Schnitzel may also be determined by joy, which is to a large degree sourced from (mis)interpreted emotional state of the other agent, and by (wrong) recognition of intended sharing by the other agent.

The fragility of cooperative behavior is also demonstrated in the high number of impact factors of the agent's emotional state. For instance, attributing emotion to perceived agents, based on their bodily expression, influence the agent's emotional state. Additionally, the perceiving agent uses this attributed emotion, together with other perceptive information (movement, valuation of the agent) to interpret the other agent's intention. In SC 2.3 we see that these factors can make a difference in co-determining cooperative behavior. The described determinants are increased (anger and its defense into guilt in case of agent A, and joy in case of agent B) by the resulting emotional transfer (triggered by perceiving the other agent's bodily expressions). Intention recognition provides an additional matching factor for images. That is, this additional functionality increases possibilities of adaptation by enabling activation of different images based on the other agent's intention. Interpreting the other agent's perceived body expression as joy and based on that - together with his movement - interpreting his intention as sharing, activates images of sharing in the agent (see Fig. 6.19). Hence, in the same personality by which sharing is

determined by defended anger to guilt, sharing may also be determined by joy and an interpreted sharing intention.

But intention interpretation not only directly co-determines behavior by activating images, it also indirectly co-determines behavior by influencing the agent's emotional state. In SC scenario 2.3.6 (see Fig. 6.20) agent A interprets agent B's body expression and movements towards the Schnitzel as intention to eat the Schnitzel alone, which increases agent A's anger and triggers the defense mechanism reversal of affect, which defends anger to guilt (as described above) in turn activating the goal to share. Agent B shares due to joyful interpretation of A's body state and cooperative intention interpretation (see bottom figure in Fig. 6.20). As seen in the bottom figure in Fig. 6.20, in agent B also anxiety increases (starting with simulation step 100) since he interprets the vicinity of agent A to the Schnitzel as ownership, which increases his unpleasure and hence anxiety and anger. However, this does not impact his decision, i.e. the joy and activation of the share-goal is (in relation) still high enough.

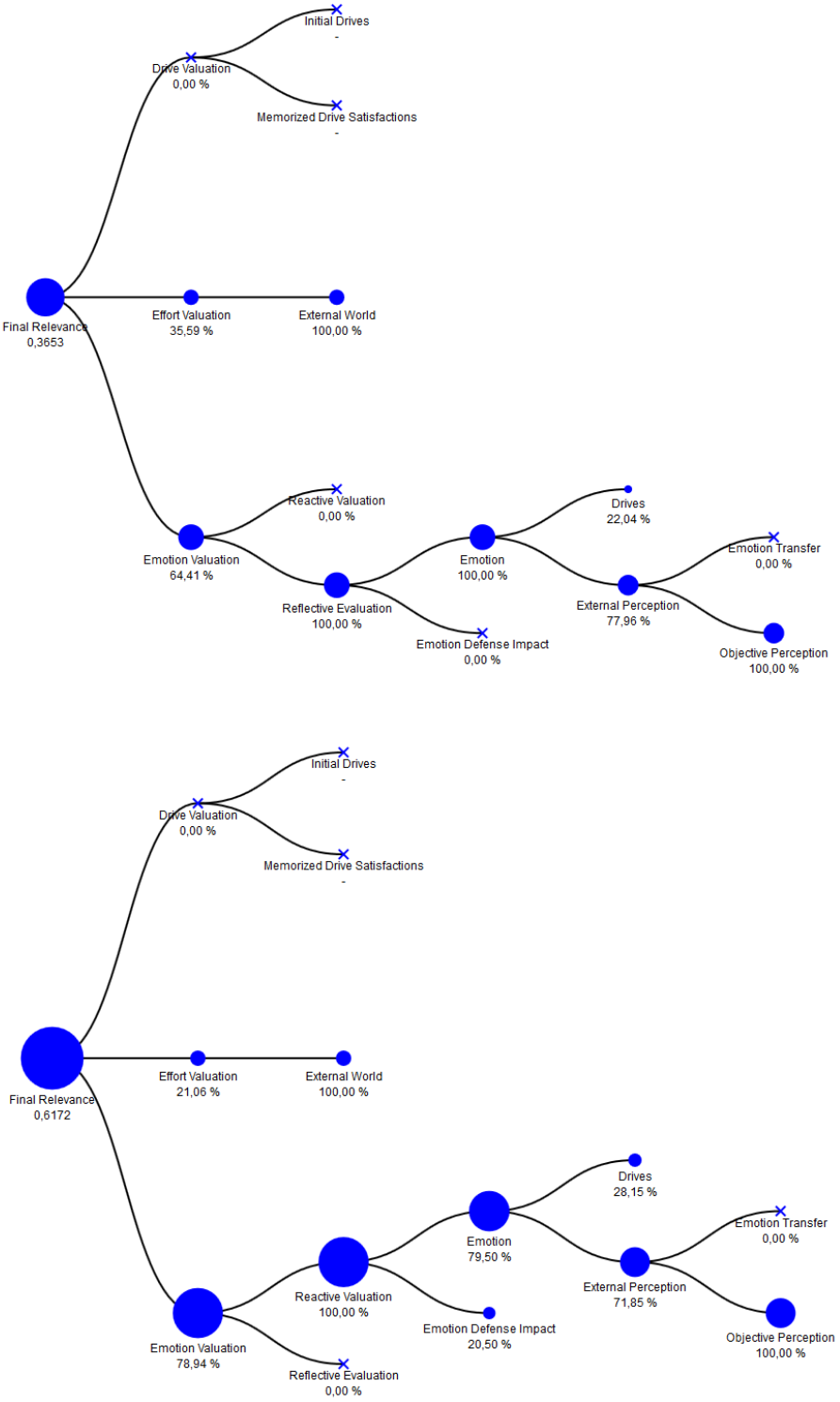

Figure 6.15: Comparison of determinant values between the reactive (bottom figure) and reflective (top figure) leave scenario. Due to the higher neutralized intensity, the agent uses his emotion as feelings to reflect on the expected benefits of the goal (to leave) in relation to his current state. Due to the stricter super-ego, extended emotion is generated in the reactive scenario, determining the agent to leave.

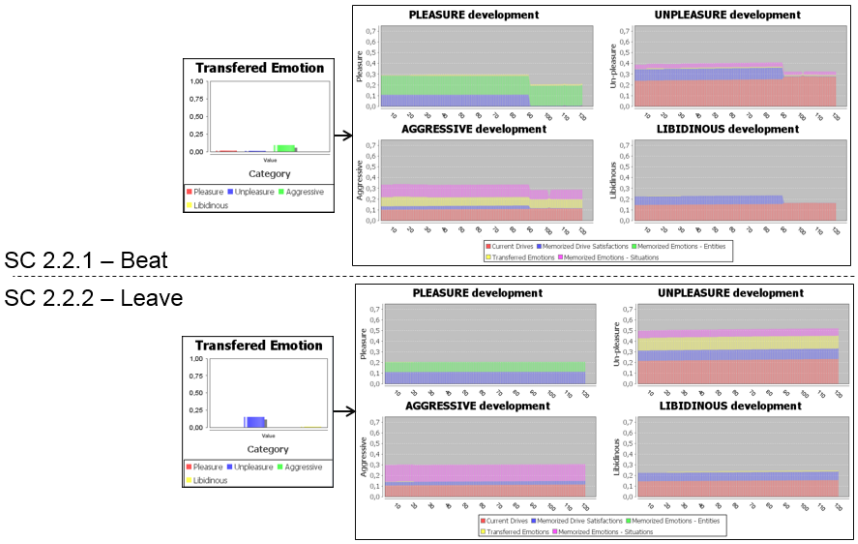

Figure 6.16: Emotion transfer as a 'tipping point' can lead the SiMA agent to leave the situation instead of beating the other agent. In SC 2.2.1 the bodily expression of the other agent is associated with anger (aggressive unpleasure in top left figure), influencing the agent's aggressive emotion component (yellow stripe in top right figure). Since the agent uses knowledge about its own state, co-determining the agent to beat the other one. In SC 2.2.2 the bodily expression of the other agent is associated with anxiety, which in this case corresponds to the subjective emotional state of the other agent, co-determining the agent to leave the situation.

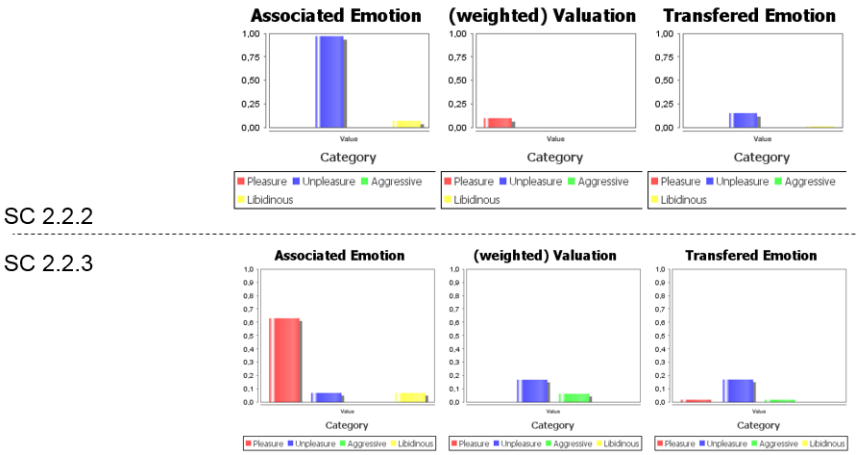

Figure 6.17: Two ways of increasing anxiety by emotion transfer in the leave-scenario: In SC 2.2.2 the bodily expression of the other agent is associated with anxiety and the other agent is valuated positively. In SC 2.2.3 the bodily expression of the other agent is associated with joy and the other agent is valuated negatively.

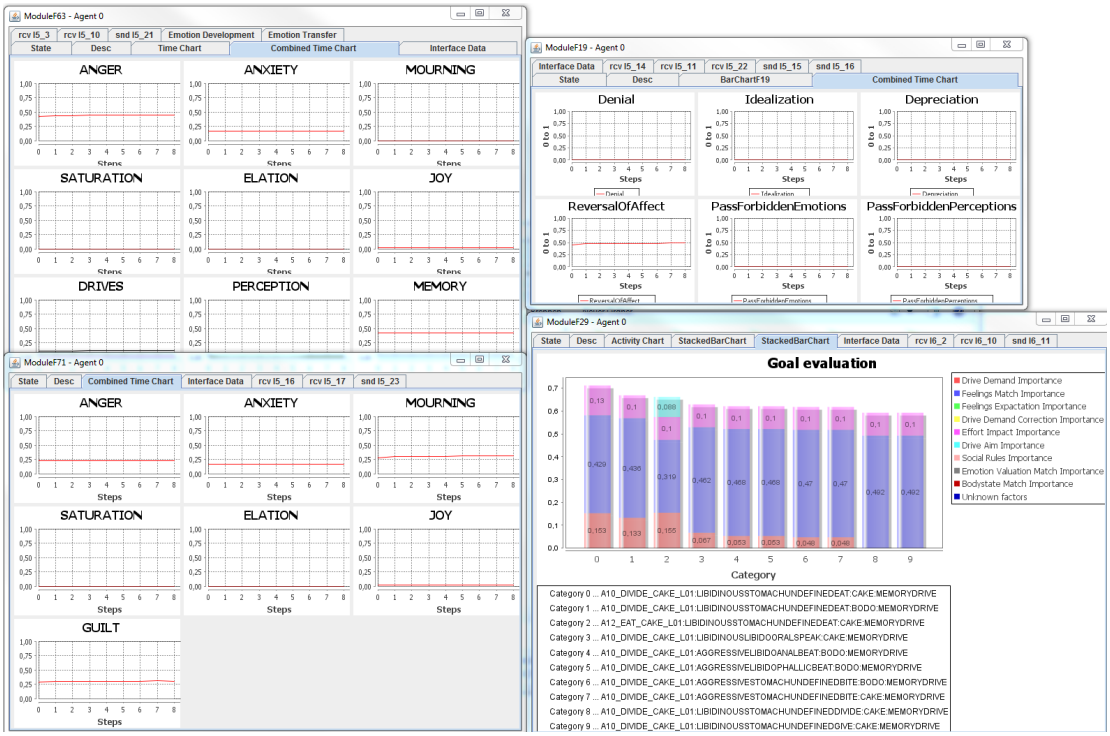

Figure 6.18: Simulation sequence of the share scenario 2.1.5. High anger can lead to cooperative behavior, if the agent has a strong super-ego and the defense mechanism reversal of affect (see top right figure) that defend anger to mourning and guilt (see left figures) by generating a conflict and displacing values from aggressive to libidinous emotion component.

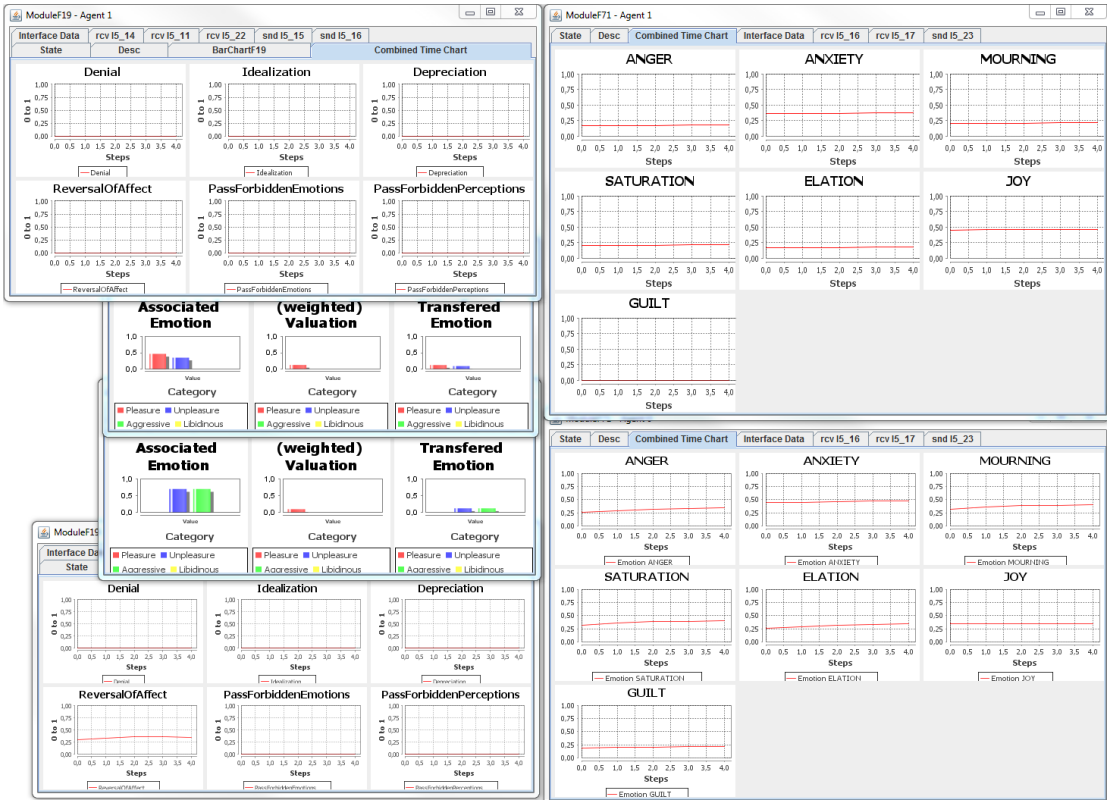

Figure 6.19: Simulation of the SC scenario 2.3.5. - share. Emotion-based defense mechanisms can co-determine cooperative behavior (defense mechanism 'reversal of affect' in the green agent, see bottom left figure), but are not necessary for cooperative behavior (no defense mechanisms are active for the red agent, see in top left figure). A positive interpretation of others' body expression (cheek redness) is sufficient to change the agent's behavior from beating (SC scenario 2.3.4.) to sharing (SC scenario 2.3.5.). This happens due to (1) increasing joy via emotion transfer and (2) interpreting the others joyful movement towards him as intention to share.

Figure 6.20: SC scenario 2.3.6. Both agents share due to different determinants and mechanisms. Agent A (top figure) shares due to guilt (which is defended from anger) and Agent B (bottom figure) shares due to emotion transfer of joy and interpreting agent's A intention as sharing (which activates the goal to share in agent B).

We see in SC 2.3 that also the spatial configuration co-determines cooperative behavior by providing additional matching factors for memory retrieval: Changing the spatial situation (Schnitzel is not in the middle between the agents, but behind the red agent, see 6.21) indirectly co-determines cooperative behavior, since the question of ownership comes to the fore. In this case a request action is executed (if the agent is a libidinous type and currently has high neutralized intensity, see SC scenario 2.3.2) by the green agent. If the red agent associates the perceived request-action with anger, he denies sharing and eats by himself; in case of associating it with guilt, he gives the Schnitzel to the green agent (see Fig. 6.22).

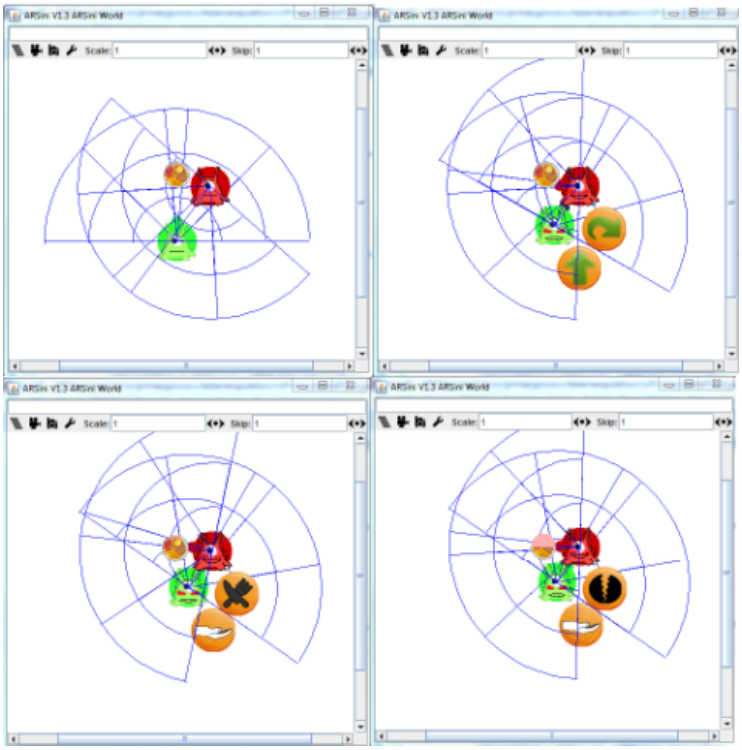

Figure 6.21: SC scenario 2.3.2. Request - share.

6.1.4 Summarizing Comparison

Figure 6.23 gives an overview of the previously described scenarios. We see that behavior is differently influenced by the three main model variables (valuations from drives, emotion/feelings, and effort) that determine the agent action (first column in Fig. 6.23). Eating is mainly determined by drives - directly from drive valuation and indirectly by drive influencing emotion (second column in Fig. 6.23). Beating is mainly impacted by the agent's emotion. Of course, not every behavior is influenced by all three main variables: Leave only gets determined by emotion valuation. We see in Fig. 6.23 how reflective use of emotion as feelings (SC scenario 2.1.2.) require less impact from feelings than in case of reactive use of emotion (SC scenario 2.1.3). The higher impact of effort valuation also shows the role of unpleasure avoidance and long-term consequences in reflective thinking. The results of alternative leave scenario 2.1.3 show how reflective thinking can be compensated by a strict super-ego, defending emotion (see last column in Fig. 6.23).

Comparison of the impact of variables in Fig. 6.23 also shows that the principle composition of emotion from drives and memorized perception is very similar. In all scenarios the impact

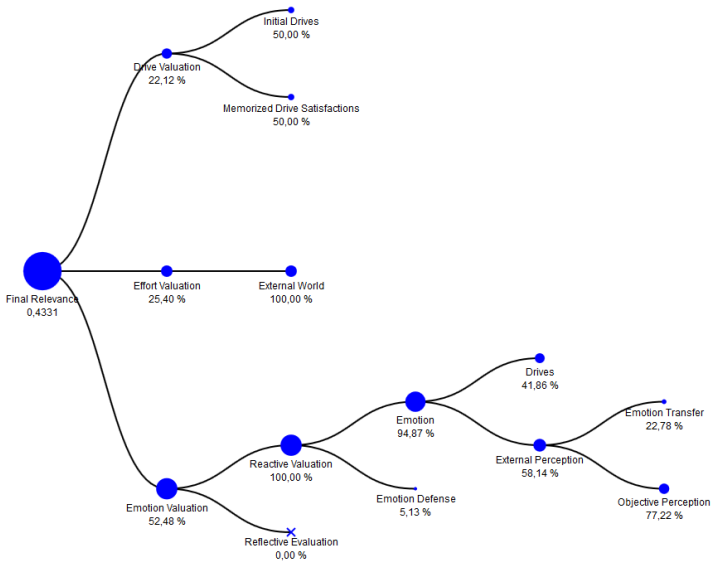

Figure 6.22: SC scenario 2.3.5. Emotion transfer can be the tipping point for sharing.

of perception on emotion is around two third (66%-77%), with the rest coming from drives. However, the variation of emotion comes from the variation in drives (aggressive or libidinous) and memorized emotion (different emotion factors).

The comparison of the model variable's impact in Fig. 6.23 also gives an overview why a change of determinants has led to a change in behavior.

Emotion transfer suffices to get the agent to beat the other instead of eating the Schnitzel (SC scenario 2.2.0 to 2.2.1). By increasing the emotion reaction sensitivity together with the positive valuation of the other agent, emotion gets a higher impact in decision making and drives less impact, respectively (see first column). This higher impact is provided by a rise of emotion from attributing emotion to the other agent and get influenced by it (third column).

The simulation cases explore different ways to get the agent to leave the situation instead of beating. One is by harnessing reflective thinking (SC scenario 2.1.2), i.e., using emotion as feelings and relates the expected pleasure and unpleasure to the agent current pleasure and unpleasure. This is reached by increasing neutralized intensity, leading to reason about the goal's consequences on the agent's current state and the effort of its execution (higher impact in first column). Another way of getting the agent to leave the situation instead to beat the other agent is by emphasizing social norms, i.e, increasing super-ego strength, shown in scenario 2.1.3. We see that emotion (anger) is defended (last column in Fig. 6.23, see also Fig. 6.14).

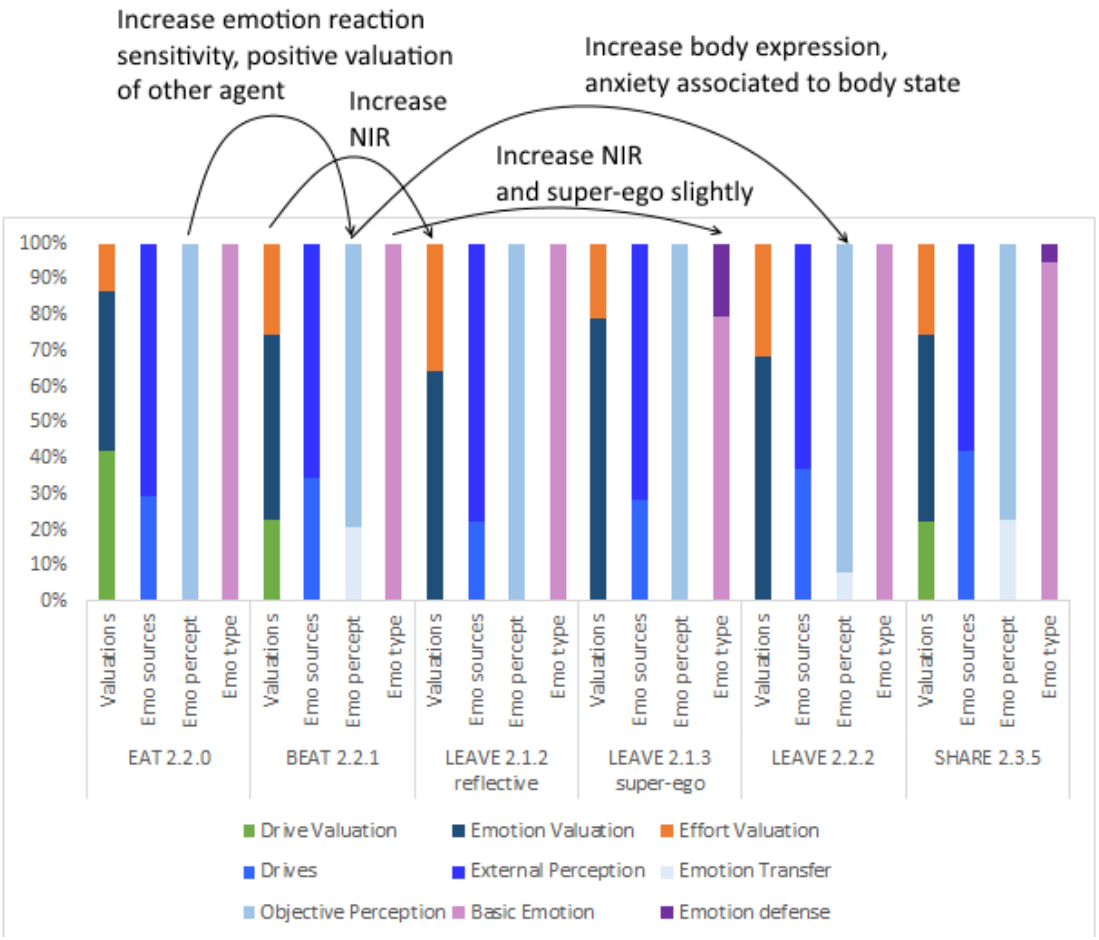

Figure 6.23: Impact of the main model variables in selected simulation scenarios. The agent's decision is mainly determined by valuations from drives, emotion and effort (first column). Emotion is sourced in drives and external perception (second column), which is influenced by emotion transfer from attributing emotion to perceived agents and by memorized emotions activated by perceived objects, including agents (third column). Emotion may be influenced by defense mechanisms (fourth column). NIR stands for the personality parameter neutralized intensity ratio, regulating the different processing modes (primary and secondary process) in the mental architecture.

Figure 6.24: Comparison of simulation case scenarios 2.3.4 (top) and 2.3.5 (bottom). Emotion transfer is able to determine sharing. Changing the memorized association to blushing from anger to joy suffices (in this determinant constellation) to get the agent to share the Schnitzel instead of beating the other agent. The change of the emotional state (left big figures) is reached by emotion transfer (left small figures). The right figures show how in the share scenario emotion transfer impacts the pleasure component of emotion(yellow stripe).

The simulation cases also explore the different ways to determine sharing. One way to get the agent to share instead of eat (scenario 2.1.0 to 2.1.5) is to change the drive type from aggressive to libidinous and having angry memories about the other agent. That is, in combination with super-ego, anger is required to defend it into guilt, which is the main determinant for sharing in the simulation case scenarios. As shown in the last column of the share action (see Fig. 6.23) defended emotion has a high impact on the current emotion and in turn on the agent's decision. The necessary anger can also be reached by emotion transfer and intention attribution (see third column in scenario 2.3.5.). Here the simulation case showed how sometimes a change of the emotional association of the other's body expressions

suffices to get the agent to share instead of beating the other (SC scenario 2.3.4 to 2.3.5). To see the details of how emotion transfer led the agent to share instead to beat, Fig. 6.24 contrast the visualization (screenshots of the simulation) of the two scenarios.

6.1.5 Interdisciplinary Double Blind Experimentation

The model was validated against fine-grained expectations from psychoanalyst using simulation cases as test templates. In simulation fine-grained explanations for these expectations were provided. It must be emphasizes that the spectrum of scenarios was simulated by only changing model parameterization. For a deeper model validation and exploration, the next step is to conduct interdisciplinary 'double blind experimentation'.

To set a focus, we reduce the number of determinants we aim to manipulate, and test our beliefs regarding the determination strength of single determinants. Our choice of determinants is built on the experience we gained during the calibration of the simulation cases. Using the experience gained in the simulation case results, we chose determinants that we expect to have a strong influence on an agent's behavior. Incrementally changing the determinant's value, we observe behavior changes in simulations. Parallel and independent (i.e., without knowledge of the outcome), psychoanalysts formulate expectations about the influence of chosen determinants. In particular, they predict the required change of the determinants and the expected behavior change. After finishing the simulations, we compare the results with the previously unknown expectations of the psychoanalysts. For instance, one of the chosen question is about the required increase of the neutralized intensity ratio to change the agent's behavior from eating (given initial low neutralized intensity ratio and fixing all other determinants). The psychoanalysts get for each experiment information about (1) the fixed constellation of determinants, (2) the determinant that will be varied, and (3) the simulation outcome with the initial value of the determinants. In case of missed expectations of the psychoanalysts, the simulation results will be interpreted and provide additional feed-back-information for model development.

We chose the focus of the SC 2.2. (see Fig. 6.1) to conduct four experiments, sketched in Table 6.1. To reduce the impact of other determinants on the agent's decision making, we first defined a fixed base constellation for all experiments, and later adapt the constellation on the respective experiment - aiming for minimal interaction between the focussed determiant and the constellation's determinants. This determinant constellation is chosen to provide a stable simulation without determinants dominating the agent's decision making. The constellation includes, for example, low emotion transfer, few super ego rules and minimal defense. After specifying the base constellation, we adjusted it for the respective experiment to focus on the determinant in question - the focused determinant should be the main player of the experiment. The second column in Table 6.1 gives an overview of the adjustments of the basic constellation. The focused determinant are expected to have a direct impact on decision making, but more importantly also an indirect impact via different variables. The first column of Table 6.1 also shows the determinant's starting value and the direction of variation.

The four experiments provided valuable insight on partially unexpected aspects of the model. The goal to produce a stable constellation proved challenging, since our simulation cases mostly aimed at illustrating behavior by showing the impact of determinants via mechanisms, where a stable constellations aims at evening out the influence of all mechanisms, to avoid 'fitting' the constellation to a desired simulation outcome. The determinant constellation and variation of the focused determinants showed the limited control-ability in parametrizing the model. The complexity of the model (in particular the interdependence of the model variables) hinders a straight-forward parameterization. New parameterizations often raise inconsistencies on the implementation and model level. This indicates that for a model of this scale, we need to test the model in a wide range of possible parameter-combinations to get a stable, easy to parametrize model.

Experimentation Determinant (value range)	**Constellation**	**Outcome**
NI-ratio (low - high)	High Drive Tensions	Eat
Super-Ego-Strength (high - low)		Eat
Bodily Expression (anger- nothing)	Medium emotion transfer	Give
Memorized Agent Valuation (negative to positive)	Slightly reduced drive influence	Leave

Table 6.1: Setup of double blind experimentation (SKZ+15)

Experiment	**Simulation Outcome**	**Psychoanalytic Prediction**
NI-ratio	'Sharing' at very high NI-ratio	'Sharing' at low NI-ratio
Super-Ego-Strength	'Giving' at low Super-Ego-Strength	'Sharing' on mid-high to high Super-Ego-Strength
Bodily Expression	No change in behavior	'Eating' on low anger
Memorized Agent Valuation	'Eating' at very low negative valuation	'Beating' on low positive valuation

Table 6.2: Experimentation results (SKZ+15)

In summary, we demonstrate the feed back from the experiments with the according levels of our methodology (see Chapter 2 and 2.2). See also Table 6.2 for a summary. The NI-ratio experiment (#1) was successful in simulating the psychoanalytic prediction, but revealed implementation problems that were not covered by our standard software tests. This showed the relevance of increasing software testing. The Super-Ego-Strength experiment (#2) produced a slightly different outcome than the psychoanalytic prediction. This results has shown the necessity for determinant adaption, which showed again the problems in mapping the qualitative values (e.g 'low') used by psychoanalyst to the quantitative values (e.g. 0.3) used for a computational simulation. It is not surprising that this happened with two similar actions as sharing and giving (see Table 6.2). The lack of behavior change in the Bodily Expression experiment (#3) pointed to a model problem. Further investigations with the psychoanalysts showed that our model did not sufficiently consider the impact of an agent's drive state on emotions attribution to others. Experiment #4 showed a clear deviation of the simulation behavior and psychoanalytic prediction. Analysis of the experiment showed a faulty configuration. The memorized agent valuation, which was varied during the experiment, is strongly related to several other parameters that should have been adjusted in correlation with the memorized valuation. Failing to do so created inconsistent parameterizations that caused faulty results. Enhancing our test methodology to prevent such mistakes are part of our future work.

6.2 Social Media Simulation of Environmentally Friendly Decisions

As described in chapters 1.2.3 and 4.1, the general question of the thesis at hand is approached by two exemplary cases ('Two agents and a food source' and 'Social Prompting'). After simulations

of cooperative behavior, simulations of environmental decisions are described. As mentioned in Chapter 5 the specific questions behind the exemplary cases pose different requirements for modeling and simulation and thus are described separately.

The simulations of basic behavior and social interactions (eat, beat, leave, share, give) showed that basic psychoanalytical concepts and Damasio's neuroscientific concepts can be implemented and simulated. In particular, the ability to transform them into a computational model showed their consistency. The ability of meeting the expectations from the original psychoanalytical and neuroscientific theories in computer simulations corroborated the underlying theories.

After this basic validation, we transform the same concepts into a more unified and dense model (see Chapter 5) to meet the requirements of the exemplary case for Social Prompting. The resulting SiMA-C model is also better able for the next level of validation using empirical data. Therefore, bridging the model parameters as well as the survey questions is necessary. After parametrizing the model according to empirical data, we are able to test if and how the SiMA-C model replicates observable reality. Although empirical validation is the focus in this chapter, it is not a means in its end, but enables us to go beyond the abilities of empirical tests. In this regard simulation cases are used. Hence, different than before, simulation cases are used for exploration, since the validation is done by comparison to empirical data.

6.2.1 Social Prompting: Think about Environmentally Friendly Actions! Conform by Following Your Own Norms

To be able to identify the variables and questions required for an empirical survey, firstly a simulation scenario is developed based on the social media exemplary case in Chapter 4.1. Secondly, survey data according to the identified variables are gathered (using first model assumptions and the specification of the exemplary case). Thirdly, the simulation model is used to replicate the survey and empirical experiments. After such empirical validation of the parametrized SiMA-C model, simulation case scenarios are specified to explore causal relations of variables for the sake of getting insights of how to motivate people in different internal and external situations to behave environmentally friendly. Hence, this step enables to go beyond the possibilities of empirical experimentation.

6.2.1.1 Bridging Empirical Experimentation and Simulation

As shown before, an exemplary case is used to specify the research question at hand and support requirements analysis and modeling. Therefore, a simple scenario of social prompting is described (see Chapter 4.1.2). However, different than before we mainly use it as an instrument for empirical parametrization and subsequently for model exploration in simulations.

The specification of required variables, which should be questioned in a survey, is supported by a dense description of our basic assumptions (see below and (SDG$^+$15)) based on the concepts described in Chapter 4.3. This process is corroborated by a literature review conducted in cooperation with psychologists from the Institute of Marketing & Consumer Research at the Vienna University of Economics and Business, which additionally strengthen the justification of choosing these variables. A review by a psychoanalyst in the SiMA-team at TU Wien (Roman Widholm) and a psychologist of the WU Wien (Stephan Dickert) established the variables' plausibility. Overall, the assumptions state that norms and emotions are central factors in decisions that influence and are influenced by a person's social context.

1. All decisions are based on subjective valuations. Affective valuations are the primary determinants (i.e., antecedent) of decision making.
   (a) Cost-benefit analysis as a secondary decision factor is done based on valuations.
   (b) Valuations primarily signal approach-avoidance behavior, which can be represented as pleasure/unpleasure valuations.
   (c) By parametrizing drives and emotion (i.e., change their influencing factors), a change in decision making can be reached.
   (d) Positive emotions (dominance of pleasure) are triggers of reinforcement as well as beacons for choices. Negative emotions (dominance of unpleasure) are triggers of motivated change as well as beacons for choices.
2. Social norms
   (a) Social norms operate via affective valuations.
   (b) Social norms have an effect through causing and prolonging conflicts.
   (c) Conforming to social norms will generate positive emotion. Not conforming to the social norms will generate negative emotion.
   (d) Descriptive norms, whose activation is dependent on the perception of norm conformity, and injunctive norms, who exist independently from other's norm conformity, are distinguished.
   (e) Group conformity is represented through agent-specific social norms. The relevance of conformity therefore depends in the personality of the agent as much, as it is affected by the situational activation through the surrounding or internal states.

These assumptions are then exemplified into a process description, based on the exemplary case. The input parameters (for a simplified sketch see Figure 6.25) of this process is then used to formulate the questions of a survey for enabling a direct empirical parametrization with minimal need to map the data. In the same time, missing input parameters of the process description are identified and necessary model adaptions are conducted. As a result model parameters and according survey questions are specified (see Table 6.2.1.1).

The survey on green electricity, which was designed and implemented[3] to explore the reasons for choosing green energy, consisted of a questionnaire based on the previously identified variables (using the simulation case approach) about the general attitude towards the environment, social norms and personality indicators of the participants themselves. Additionally, we included a brief experiment of the awareness of social media as a suitable communication channel about switching from grey to green energy (SJD16). To identify empirical evidence for decision factors of agents, a number of characteristics and attitudes were defined and tested in the survey. To give an overview, based on the required parameters for the SiMA-C model, the following seven categories of possible influencing factors were formed: (1) Personality factor (i.e., intuitive vs. reflexive decision-making), mapped to personality parameters in the SiMA-C agent; (2) initial internal state (e.g., bodily needs as happy, sad, hungry, and the emotional state), mapped to the agents' state indicator; (3) previously activated data (i.e., daily plans and relevant activities), mapped to the agents' initial activation values of memories; (4) attitude, trust and prior knowledge

[3]The survey was designed by this thesis' author in cooperation with psychologists from the Institute of Marketing & Consumer Research at the Vienna University of Economics and Business, who conducted the survey.

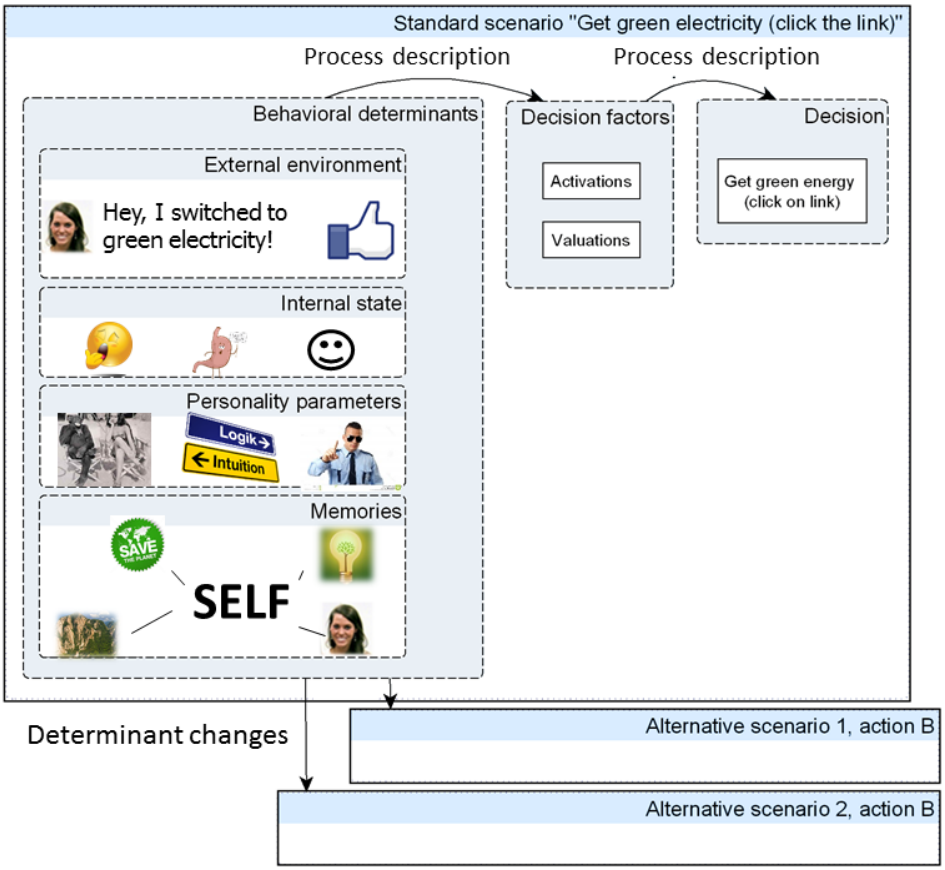

Figure 6.25: Towards a simulation case specification: Simplified sketch of the identified input parameters for the 'Social Prompting' exemplary case. The interaction of the identified parameters in a model of activation and valuation is expected to determine environmental friendly behavior in the agent. These are: the message from a social-media-friend about how good it is to switch to 'green electricity'; parameters about the super-ego-strength (internalized social norms), imitation type, and intuitive tpye; memories about the valuations of nature, green electricity, environmental norms, and the social-media friend; parameters of the current internal state (bodily needs, emotional state).

about green energy, mapped into valuations of green energy, (5) emotions related to green energy (e.g., pride, envy, guilt), mapped to associated memories, confidence, and summary valuation of green energy; (6) social norms related to environmental behavior (i.e., descriptive and injunctive norms), mapped to memorized norms and their strength and implemented by different ways in activating norms; and (7) self-efficacy in relation to environmental protection, mapped to the agents' personality parameters.

Survey variables	Model variables
Intuitive and reflexive decision making: e.g. Most of the time I trust my gut feelings. e.g. Before making a decision I weight pro and cons.	Personality factor neutralized intensity ratio
Current emotional state e.g. How anxious/sad/happy/motivated/stressed are you?	Initial pleasure, unpleasure, and conflict values
Current goals e.g. How relevant is work/shopping/relaxing... right now for you?	Initial activation value of goals
Attitude e.g. How positive/negative is your attitude towards green electricity? e.g. Green electricity (would) satisfy me.	Memorized valuation value of green electricity
Involvement e.g. I know about the (dis)advantages of green electricity e.g. I have gathered much information about green electricity I know the difference between green and grey electricity and difference in their valuation	Memorized valuation and initial activation value of green electricity Association weight between green and grey electricity
Trust e.g. It is not guaranteed that green electricity is only coming from sustainable sources.	Confidence of memorized valuations
Social norms e.g. I would feel proud (guilty) when I (do not) protect environment.	Memorized valuation of norm to protect environment.
Social norms: guilt e.g. I would feel guilty if I would not switch to green electricity after my peers did.	Personality factor imitation sensivity
Social norms: envy e.g. I would feel envious if peers have green electricity	Personality factor ownership conflict
Injunctive norms: It would be the right decision to switch to green electricity	Memorized valuation of goal to get green electricity
Ease to switch e.g. I suppose switching to green electricity is easy	Memorized effort-unpleasure in goal to get green electricity

6.2.2 Empirical Validation

After a successful mapping from model parameters and variables to survey questions and vice versa, the result of a survey analysis is mapped into a model parametrization to examine the replication of empirical data in simulations[4].

Survey Analysis (SJD16)

Regression analyses of the survey data reveal that the perceived ease of switching energy providers, personal involvement, and higher descriptive norms (i.e., how many people in one's social environment already use green energy) explain 43% of the variance in participants' current energy choice (i.e., which energy source they currently use). These predictors correctly identify 75% of green and 90% of grey energy consumers. Additionally, people's intentions to switch from grey to green energy can be predicted by participants' attitude towards green energy (i.e., more favorable attitudes are connected with stronger intentions to switch), personal involvement, and injunctive norms (i.e., whether one should switch). These three predictors can explain 28% of the variance in intentions to switch from grey to green energy, and they correctly forecast 86.4% of those people who are willing to switch and 57.3% of those who are not willing. Finally, these variables as well as emotion variables (such as pride, envy, and guilt) were used to create different clusters of energy consumers, which can later be used as a blueprint for specific agent types to provide better ground for agent parametrization and empirical replication in simulation. To do so, three analyses were done with different variable combinations. In the first analysis the variables ease of switching, involvement and descriptive norms were used to form the following four homogenous clusters on base of the current energy resource: (Cluster 1) Grey energy consumer with high involvement, (Cluster 2) Grey energy consumer with low involvement, (Cluster 3) Green energy consumer with low descriptive norms, and (Cluster 4) Green energy consumer with high descriptive norms. In the second analysis, the variables attitude, involvement, injunctive norms, pride, envy and guilt were used to form four homogenous clusters on basis of the willingness to switch: (Cluster 5) High willingness to switch with weak emotions, (Cluster 6) High willingness to switch with strong emotions and high descriptive norms, (Cluster 7) Mixed group with weak emotions, and (Cluster 8) High willingness to switch with strong emotions and low descriptive norms. Finally, the third analysis used the variables attitude, involvement, injunctive norms, descriptive norms, ease of switching, pride, envy and guilt to form three homogenous clusters on base of the willingness to switch: (Cluster 9) Mixed group with weak emotions, (Cluster 10) High willingness to switch, and (Cluster 11) Low willingness to switch with low involvement.

Replication of Survey Findings in Simulation

To replicate the findings of the survey, agents were parametrized according to the clusters from the survey. We normalized the Likert scale (1-5) used in the survey to a range of 0-1 and parametrized the agents' variables according to the mapping described above. Using a social media scenario in simulations (where an agent gets a message from a friend about switching to green electricity, see Chapter 4.1.2), the goal's relevance to get green electricity is compared to the participants' tendency to get green electricity, which reflects a question in the survey (see Figure 6.27).

We achieved a sufficient replication of the empirical data without any need for model calibration (i.e., adapting the parameters to get a sufficient mapping): The tendency to get (or have) green

[4]The mapping approach using cluster analysis is developed and conducted by this thesis' author. The analysis of the survey results (see next chapter) is conducted by psychologists from the Institute of Marketing Consumer Research at the Vienna University of Economics and Business (Stephan Dickert and Erdem Geveze).

	personality factor: "psychic neutralization rate"	initial emotion: aggression	initial emotion: pleasure	initial emotion: libido	stomach drive	stamina drive	Action "watch children", preactivation	Action "watch children", valuation	Action "meet friends", preactivation	preactivation action "rest"	Action "go shopping", preactivation	Action "go to work", preactivation	Association weight: Green el <-> grey el.	pre-activation: "get (information about) green electricity"
Cluster 1	0,38	0,18	0,63	0,24	0,34	0,46	0,28	0,50	0,40	0,65	0,50	0,66	0,77	0,56
Cluster 2	0,45	0,10	0,72	0,22	0,41	0,53	0,33	0,50	0,29	0,71	0,49	0,58	0,37	0,14
Cluster 3	0,38	0,20	0,70	0,26	0,35	0,51	0,48	0,50	0,44	0,72	0,57	0,77	0,88	0,78
Cluster 4	0,41	0,20	0,68	0,31	0,33	0,48	0,36	0,50	0,46	0,68	0,50	0,73	0,63	0,51
Cluster 5	0,36	0,19	0,63	0,33	0,26	0,51	0,33	0,50	0,45	0,66	0,50	0,70	0,73	0,44
Cluster 6	0,43	0,24	0,66	0,34	0,38	0,49	0,43	0,50	0,41	0,65	0,48	0,69	0,59	0,45
Cluster 7	0,40	0,15	0,64	0,23	0,30	0,49	0,22	0,50	0,39	0,67	0,49	0,68	0,62	0,34
Cluster 8	0,37	0,20	0,70	0,25	0,43	0,52	0,31	0,50	0,41	0,76	0,56	0,68	0,81	0,68
Cluster 9	0,38	0,20	0,64	0,29	0,31	0,45	0,33	0,50	0,46	0,66	0,52	0,66	0,66	0,46
Cluster 10	0,41	0,21	0,68	0,28	0,38	0,51	0,34	0,50	0,44	0,70	0,52	0,69	0,75	0,61
Cluster 11	0,47	0,10	0,73	0,16	0,38	0,51	0,22	0,50	0,27	0,71	0,47	0,60	0,50	0,25

mem Valuation of green el.: Pleasure	mem Valuation of green el.: Unpleasure	mem Valuation from norms of get green el.: confidence (=	MEM Val. By Norms: get green el. (pl)	mem Val. By Norms: get gray el. (pl)	mem Val. By norm: protect env. (pl)	personal Factor: Conflict sensitivity	personal Factor: imitation sensitivity	personal Factor: ownership conflict sensitivity	personal Factor: super ego strengh	mem Val by norm.: conform to friends (pl)	mem val.: be active (pl)	mem val.: be active (unpl)	Memorized emotion to nature: pleasure	Memorized Emotion to nature: unpleasure	MemorizedEmotion to nature: libidous	asso. Weight: self <-> nature
0,60	0,64	0,51	0,58	0,60	0,64	0,27	0,42	0,18	0,73	0,58	0,78	0,54	0,85	0,85	0,79	0,74
0,33	0,35	0,27	0,45	0,42	0,57	0,18	0,40	0,09	0,70	0,66	0,67	0,45	0,80	0,84	0,69	0,64
0,85	0,77	0,68	0,86	0,80	0,86	0,60	0,51	0,50	0,78	0,70	0,91	0,70	0,93	0,87	0,91	0,83
0,59	0,56	0,48	0,64	0,59	0,70	0,43	0,52	0,37	0,71	0,71	0,77	0,54	0,86	0,79	0,80	0,74
0,72	0,66	0,49	0,72	0,71	0,76	0,32	0,42	0,06	0,76	0,65	0,85	0,69	0,95	0,95	0,92	0,86
0,53	0,54	0,47	0,60	0,52	0,67	0,44	0,51	0,42	0,71	0,68	0,74	0,55	0,81	0,80	0,78	0,74
0,42	0,51	0,38	0,43	0,51	0,55	0,13	0,36	0,03	0,71	0,58	0,71	0,46	0,83	0,84	0,75	0,69
0,82	0,69	0,59	0,84	0,75	0,80	0,60	0,55	0,57	0,73	0,71	0,88	0,66	0,90	0,84	0,88	0,78
0,51	0,56	0,45	0,51	0,54	0,60	0,25	0,41	0,14	0,70	0,61	0,73	0,52	0,82	0,82	0,76	0,75
0,76	0,64	0,56	0,79	0,69	0,81	0,52	0,54	0,46	0,73	0,69	0,84	0,61	0,90	0,82	0,86	0,77
0,25	0,41	0,32	0,28	0,34	0,42	0,04	0,34	0,01	0,72	0,60	0,66	0,37	0,80	0,85	0,69	0,59

Figure 6.26: Paramterizing agents accordding to clusters from the survey analysis.

electricity can be predicted by the SiMA-C model. However, analyzing the relevant variables in the clusters that are not replicated very well (e.g., cluster 2 and 11), we assume that the interplay of valuation and evaluation (see Chapter 5.6) does not sufficiently cover the interplay of intuitive and reflective decision. This seems especially the case for considering variables of involvement. This interpretation provides input for future model development. However, putting the non-goal of a detailed replication beside, the goal was not to replicate the exact empirical data in every detail, but to provide an explanation model for them. In particular, to replicate how a change in parameters would result in a change of behavior. This includes showing how different variables are able to determine if and why agents would switch to green electricity, when prompted to do so in a social context. In this regard the simulated SiMA-C model was successful in accounting for that.

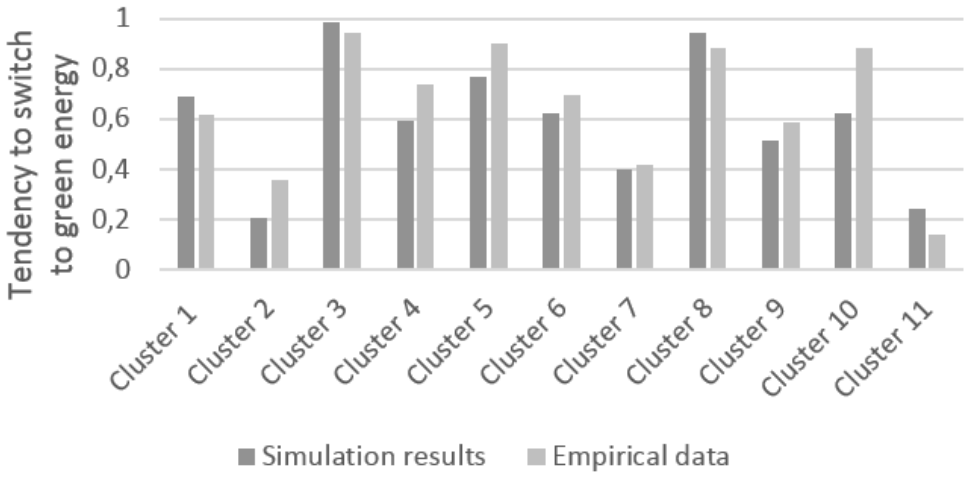

Figure 6.27: Comparison of simulation results and survey results. The SiMA-C model (dark grey) is able to replicate the tendency to switch to green electricity from empirical data (light grey) and explain the reasons (see text).

6.2.3 Model Walk-through Using a Social Media Scenario

Next we will use different visualizations (generated in simulations) for a model walk-through to demonstrate the causal chain in decision-making and how the various impact factors of environmentally friendly behavior are covered by the generic SiMA-C model. For demonstration purposes, we use the already mentioned social media scenario that considers all mentioned impact factors (see Chapter 4.1.2.) (SDG+15). Victoria is currently working and gets a social media message from her friend Caroline about switching to green electricity and how good this felt. The configuration of simulations includes creating memories and parametrizing the model. Victoria memorizes several internalized norms, e.g., to appreciate friendship, improve the world, and protect the environment. Part of the memories are also the knowledge about how she satisfies her needs, pleasureful thoughts and her relation to nature and to other people, e.g., to Caroline. Victoria's initial internal state is configured as hungry and a bit exhausted. But nevertheless Victoria is in an overall joyful state. Her conflict state is quite high (which does not exclude the joyfulness) since she has not confirmed to her own norms recently. The personality factors of Victoria in forms about her general norm conformity and how intuitive she is in general. Due to the simulation scenario, the goal to work is configured as pre-activated. Figure 6.28 gives an overview of why the goal to get green electricity is highly ranked, given the agent's current internal and external configuration.

The high relevance value in red is mainly due to valuation (light red), evaluation does not affect the overall goal relevance. Norms that are associated with green electricity are highly activated and impact the decision (green bar). Due to the effort to get green electricity mainly reality conflicts exist. Unsurprisingly the primary source of valuation are norms (in magenta) and the secondary source are direct activations from perception (blue). Figure 6.29 shows concretely how the goal to get green electricity is activated after perceiving the social media message. Perceiving the prompt from Caroline to switch to green electricity directly activates the corresponding goal. Caroline is memorized as a close friend (i.e., high association strength and positive summary valuation), which is why her normative prompt is activated highly. An additional activation comes from the perceived high number of 'Likes'. These two aspects can be described as following

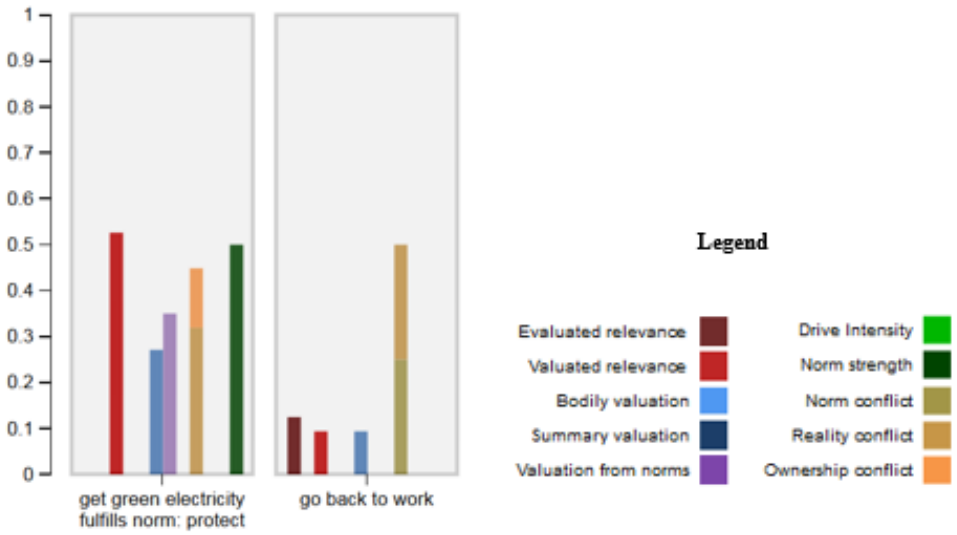

Figure 6.28: Overview of goal selection in a social prompting scenario. The get-green-electricity-goal's valuated relevance (in light red) is mainly due normative valuation (in magenta) coming from the highly activated norm (in green), but also due to summary valuation activated from perception (in blue). Evaluation from reason-based thinking (in dark red) would prefer to go back to work, but does not suffice to compete with the valuation to get green electricity.

a descriptive norm. Due to direct and indirect activation of the norm to protect environment (which is activated directly to the message and indirectly due its associations with the perceived message's background picture of nature and the prompt to get green electricity) the goal gets most of its activations from normative sources. This corresponds to following an injunctive norm. Hence, these norm types are distinguished by the type of activation. How these activations trigger the causal chain in decision-making is shown in Figure 6.29.

Due to the high activation by perception and memorized valuations perceptive and normative valuation have the highest impact. This reflects the relevance of an agent's attitude and norms (indicated by the survey), which is represented in the model by summary valuation and a configured pre-activation value.

Due to the set personality parameters of Victoria, she is more of an intuitive type, thus valuation is the main source for the goal's relevance. Additionally the low uncertainty in valuations of different goals does not require an intensive reflection on their selection. However, due to her hunger she has a high degree of neutralized intensity (which regulates dual processing, see Chapter 5) and reflects on her decision, i.e., conducts evaluation. As described, amongst others, evaluation also considers the conflict intensity of goals and consider the expected change of an agent's state. Since Victoria is initially in a conflicting state, which is increased by current conflicts due to the social media message, her normative valuations additionally impact the goal's relevance in evaluation. The current state, which is used in evaluation, has changed due to the current demands, activated data, and mediated conflicts. For instance, the high valuations of green electricity leads to an ownership conflict, since Victoria recognizes that Caroline, but not herself, has green electricity. Due to memorized association strength between Victoria and Caroline and positive summary valuations of Caroline, Victoria is able to identify with Caroline and imitate her behavior in getting green electricity. The resulting state indicator corresponds to the descriptive emotions of envy and guilt. Additionally, the evaluation process is able to consider variables of involvement

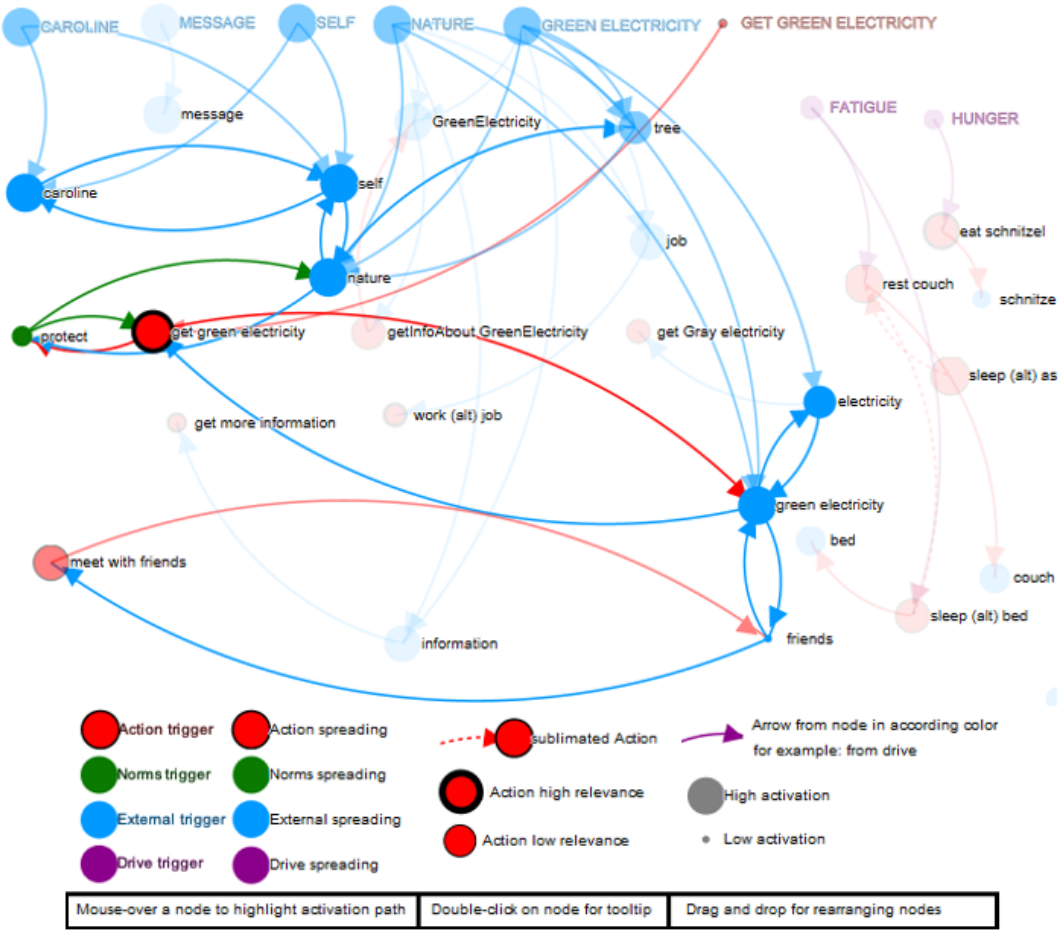

Figure 6.29: Activation pattern in associative memory for the social prompting case. The goal to get green electricity is directly and indirectly activated from perception and spreading activation. See text for description. Words in block letters indicate perceived data.

(e.g. confidence). However, in Victoria's case, the decision to get green electricity is mainly a reactive one due to activation and valuation of corresponding memories.

6.2.4 Replication of Empirical Experimentation

The survey's experimental part[5] show some indication that manipulating the visual cues given by the message's background picture influenced participants' expressed willingness to switch to green electricity (see Figure 6.31). Apparently, people are more willing to switch to green energy when they are prompted with a background picture that displays nature. However, displaying the picture has no significant impact on the willingness to recommend green energy to friends.

We further tested the impact of nature as the message's background picture in simulations and observed if the SiMA-C model is able to propose reasons for its impact on the decision. We extended the model to consider the impact of state-perception on decision-making. This functionality is triggered by perceiving an object (nature) or agent's (message-sender) state. For

[5]The survey was designed by this thesis' author in cooperation with psychologists from the Institute of Marketing & Consumer Research at the Vienna University of Economics and Business, who conducted the survey.

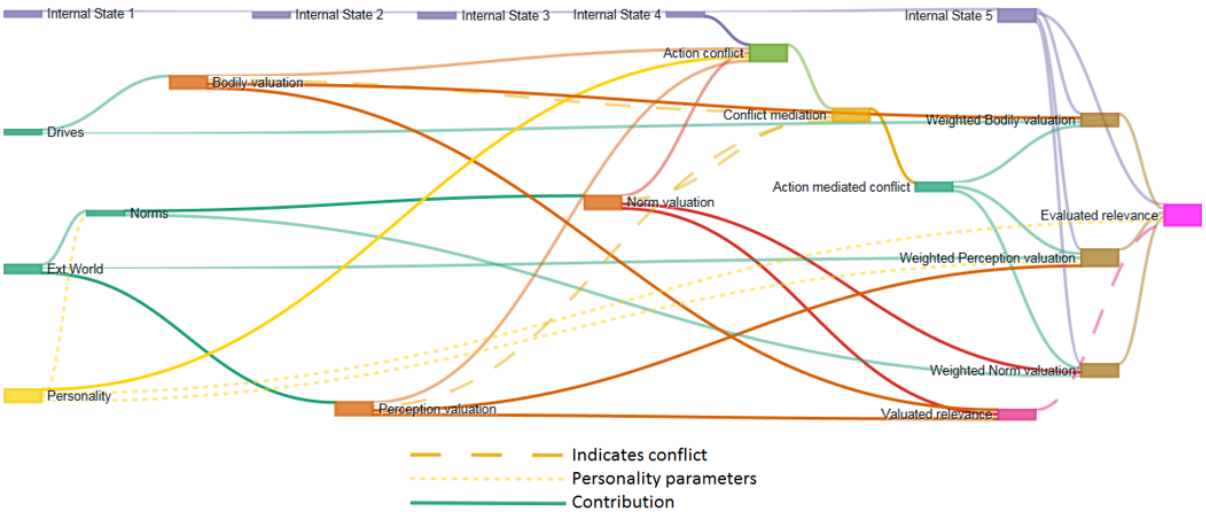

Figure 6.30: Decision evolvement of the goal to get green electricity. The four determinant groups in the left corner(internal state, drives, external world, and personality parameter) activate and valuate (in orange). Different valuation get into conflict (in yellow). Valuations get weighted (in brown) by the relevance of their activation sources (drives, external perception, norms) and related to the current internal state (in blue), i.e., the expected unpleasure avoidance and pleasure gain is related to the current unpleasure and pleasure to reason about the necessity of the goal. Based on the relevance of the different activation sources, valuations are merged to get the evaluated relevance of the goal.

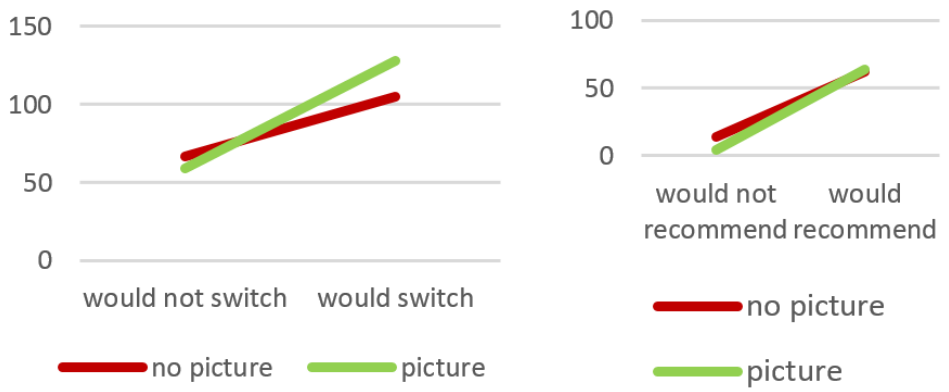

Figure 6.31: Empirical experiment: Visual cues influence willingness to switch to and recommend green energy. (Y-axes denote absolute frequencies.)

instance, another agent may indicate pleasure by corresponding bodily expressions. Based on the memorized relation and valuation of that object or the other agent (which can be parametrized according to the survey), the agent's pleasure and unpleasure state (represented by the state indicator, see Chapter 5) is altered, which influence the evaluation of goals. This process can be described as affective empathy. Similar to the survey's findings about the presence of visual cues regarding nature, simulations show an impact on the decision when perceiving a message with a picture of nature (see Figure 6.32). Additionally, the state of nature in the picture only slightly changes the tendency to switch to green energy. Hence, although it may seem counterintuitive, the state of nature in a background picture of a social media message does not seem to have an

impact on the tendency to switch to green electricity (see Figure 6.32, where an agent associates different states of displeasure and pleasure for the different pictures).

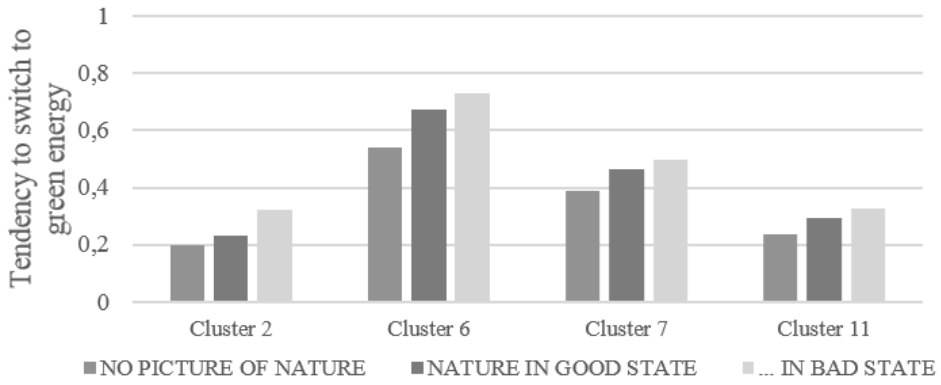

Figure 6.32: Simulation experiment: Impact of perceiving different states of nature on the tendency to switch to green energy. We used clusters with low tendency to switch to green electricity to test if a picture of nature is able to increase the tendency. Due to the different valuations of nature in the clusters, a background picture has different impact on the decision, with the state of nature generally having a low impact.

An analysis using our model provides candidate explanations: The manipulation of the agent's state (e.g., increase of displeasure due to the bad state of nature in a background picture) does not consider concrete reasons of the current state, but only the abstract sources (normative, perceptive, bodily). In other terms: Unconscious affective empathy alone is not sufficient, the agent also has to reflect on the concrete reasons of its current state to decide which goal will have the best effect on its state. That is, the message has to emphasize the relation of the agent's state to the environments state.

6.2.5 Simulation Case Specification

As mentioned, in the context of the social media scenario, simulation cases are an additional means for validation (after empirical validation) and especially for model exploration. This simulation-based method enables a flexible validation and exploration, e.g. testing the impact and interactions of variables that are not or difficult to manipulate in empirical studies.

As usual, we split the exemplary case into multiple simulation cases (see Figure 6.33), according to the model functionalities that we want to test and explore. Thereby we breake down the complexity given by the high number of variables and enabling focused development and simulation. The separation follows the areas identified in the survey (attitude, norms, involvement) and aims to extend the analysis by the possibilities of simulation. An additional area is covered by the last simulation case (empathy), exploring an area that is difficult to cover by empirical data. These areas can also be regarded as point of departures for future policy formulation. That is, they provide the basis to formulate attitude-based, norm-based, involvement-based and/or empathy-based policies to get people to switch to green electricity using social prompting. Using the possibilities given by the simulation case structure and simulation, a relevant question in this regard is which prerequisites these areas have, e.g. the dependency between involvement and confidence (tested in SC 1.3). But of course, the simulation case specification should also enable

testing the expected impact of determinants and their interactions (e.g. is one determinant able to compensate the lack of another one?).

Cluster 2 of the survey is used as a baseline scenario, since it represents an agent personality on the tipping point of switching to green electricity. On the one hand this personalities are especially relevant to address, when aiming to motivate people to environmentally friendly behavior. It would suffice (for starters) if we neglect the extreme cases and observe how a minimal change in social prompting is able to change the decision of such agents (i.e., personalities). On the other hand cluster 2 provides all variables of the mentioned areas in a medium expression, which minimizes dependencies and enable a focused exploration of the area in the single simulation cases.

The overarching topic of the simulation cases can be regarded as bridging low-level and high-level processing, individual (drives) and social (norms) aspects in decision making. As mentioned, every simulation case has its single focus in this aim.

Simulation case (SC) 1.1. (see Figure 6.33) focuses on central determinants of attitude. In the context of the SiMA-C model this concerns the dependency between activation and valuation. With the simulation case we can test the dependency of valuation from different activation source. For instance, if the main trigger of environmental behavior comes from perception, we expect that the general attitude (represented by memorized summary valuation) is key. The simulation case also aims to show that without changing the attitude (i.e., memorized valuations), agents can be motivated to environmental behavior by sufficient activation (SC scenario 1.1.3). However, it is assumed that this only can be done by increasing all activation-based parameters.

SC 1.2. looks at the norm-related determinants. Following the SiMA-C model, central mechanisms for norms to operate (besides valuations) are conflict-generation and mediation. Additionally, the simulation case introduces the differentiation between injunctive and descriptive norms and provide an explanation for their different relevance (see survey result): different types of activating the same memorized norms. SC scenario 1.2.2 (and following) demonstrates the prerequisites of descriptive norm activation, i.e., the high dependency on other determinants for descriptive norm activation to have an impact.

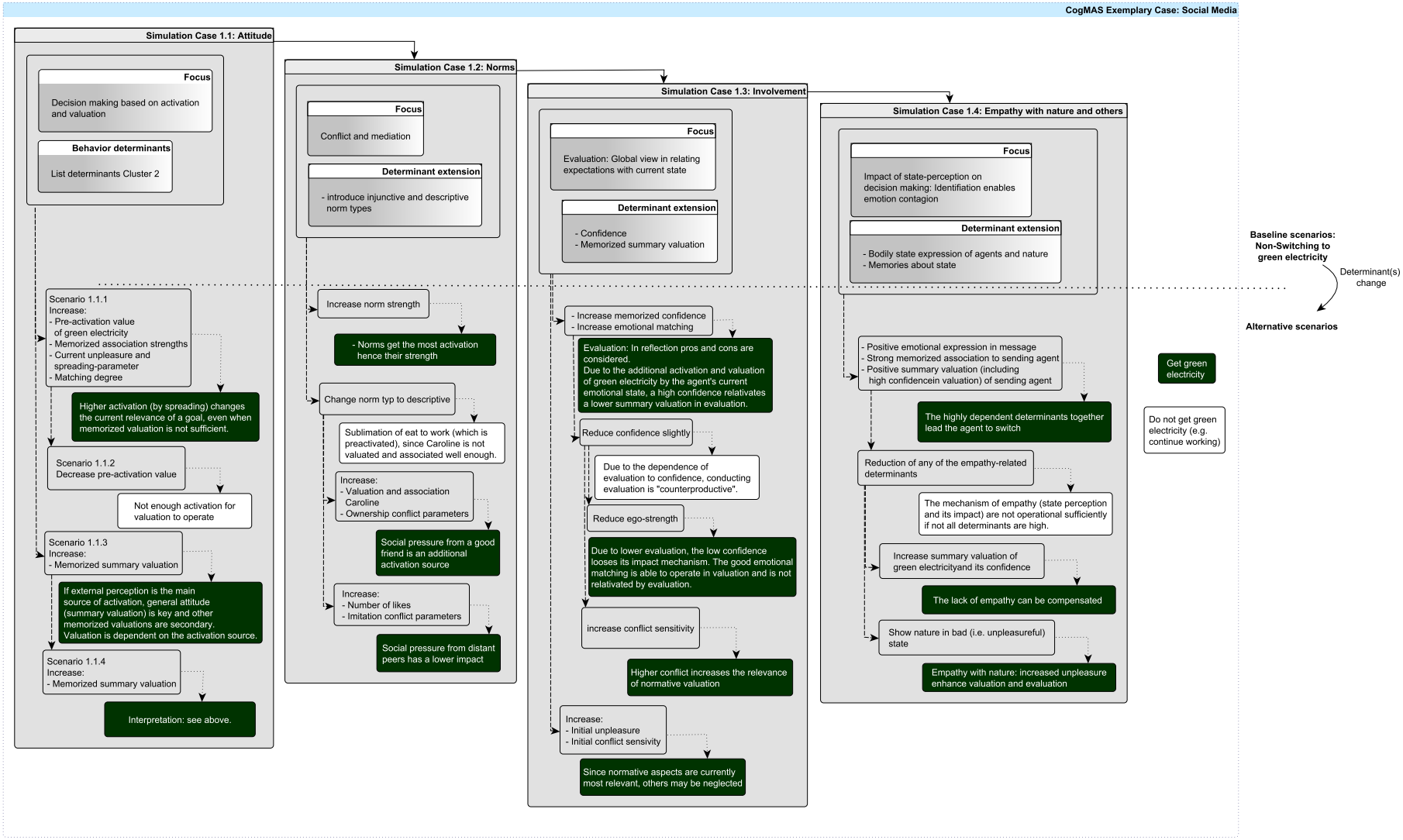

Figure 6.33: Incremental simulation cases for development and evaluation of EC 'Social Prompting'. Since a first validation is reached by replicating empirical data, simulation cases are used to explore the model and conduct simulation experiments beyond the possibilities of empirical experiments in surveys. Cluster 2 as a quite neutral agent type on the tipping point to switch is chosen for a baseline scenario. Alternative scenarios specify how a change of these determinants (transparent boxes) is expected to change the agent's behavior and specify the reasons (colored boxes).

In SC 1.3. the relation between valuation and evaluation is in the fore. Until now , the simulation cases focused on how to change the behavior of cluster 2 agents by means of the unconscious, i.e. activation, valuation, and mediation. SC 1.3. aims to explore the dependency of evaluation from determinants. This is done in an exemplary fashion using confidence in memorized valuation of green electricity. The dependency of confidence on the summary valuation lead to following assumption: Memories of green electricity have to be activated by perception and the agent's current emotional state for confidence (and subsequently evaluation) to have an impact. Overall, it is assumed that adressing determinants of evaluation without considering confidence as a prerequisite will lead to weaken the tendency to choose green electricity. Only when dual processing grade is reduced in favor of valuation, the downsides of using evaluation without considering its prerequisite is avoided (reduction of ego-strength, see scenario 1.3.3).

SC 1.4. shows that empathy (i.e., perceiving the state of other agents and/or nature and react on that) is able to have an impact, but only when all related determinants have a high value and are addressed by the social media message. These are: the agent's emotional expression (reflected in the message), confidence in the memorized summary valuation of green electricity, high association (i.e., good relation) between the sending and receiving agents, and positive valuation of the sending agent. It is assumed that a reduction of any of these determinants would not lead the agent to choose green electricity. However, it is also assumed that their reduction can be compensated by increasing other variables.

6.2.6 Beyond Empirical Validation and Experimentation

The replication of empirical data demonstrates that the SiMA-C model is a valid explanation model of environmentally friendly decision making. Next, the SiMA-C model is applied to explore the impact and interplay of determinants to develop policies for environmentally friendly behavior. As mentioned, this exploration is guided by simulation cases and testing the underlying assumptions. Hence, besides further corroboration of the model, the aim is to have an exemplary exploration of how the impact-areas identified in the survey (attitude, norms, involvement) influence unsure agents (see 6.2.5).

In the remainder of this chapter we focus on the insights provided by simulation cases 1.3 and 1.4, since the empirical studies were already clear in showing the relevance of attitude and norms, but not able to explore the dependencies in using involvement and empathy.

6.2.6.1 (E)valuating Green Electricity

An important line of explaining agents' decision using the SiMA-C model is given by valuation and evaluation. SC (simulation case) 1.1 focused on the activation and valuation part of the model. We saw that if a policy (for motivating people to switch to green electricity) would choose the means of memory activation to get people to decide environmentally friendly, it has to address many variables intensively. Also, addressing valuation requires only small changes of few variables. However, in general an external policy is rather able to address memory activation than memory valuation.

We could reason that addressing environmentally friendly behavior by means of evaluation could be more efficient. However, on the one hand evaluation is based on the result of activation and valuation, and is only extended by additional variables, hence the complexity increases. On the

other hand, we can use the global view of the evaluation mechanism in merging all variables that are relevant in the current internal and external situation. For instance, for evaluation emotion can be used and manipulated in its various forms, since a key purpose of evaluation is to relate expectations (i.e., valuation) to the agent's current state.

In any case the dependency between the variables cannot be analyzed a priori. Hence, as an example we explore confidence as a assumed key variable in evaluation (and which is not used in activation or valuation) and its relation to different forms of emotion (memorized summary valuation and the agent's current state). Exploring evaluation and relating it to the variables of valuation could help us to determine in which variable context to address mechanisms of valuation or evaluation, respectively. The simulation results underline these assumptions and show how the resulting decision is indeed achievable by manipulating evaluation-based variables.

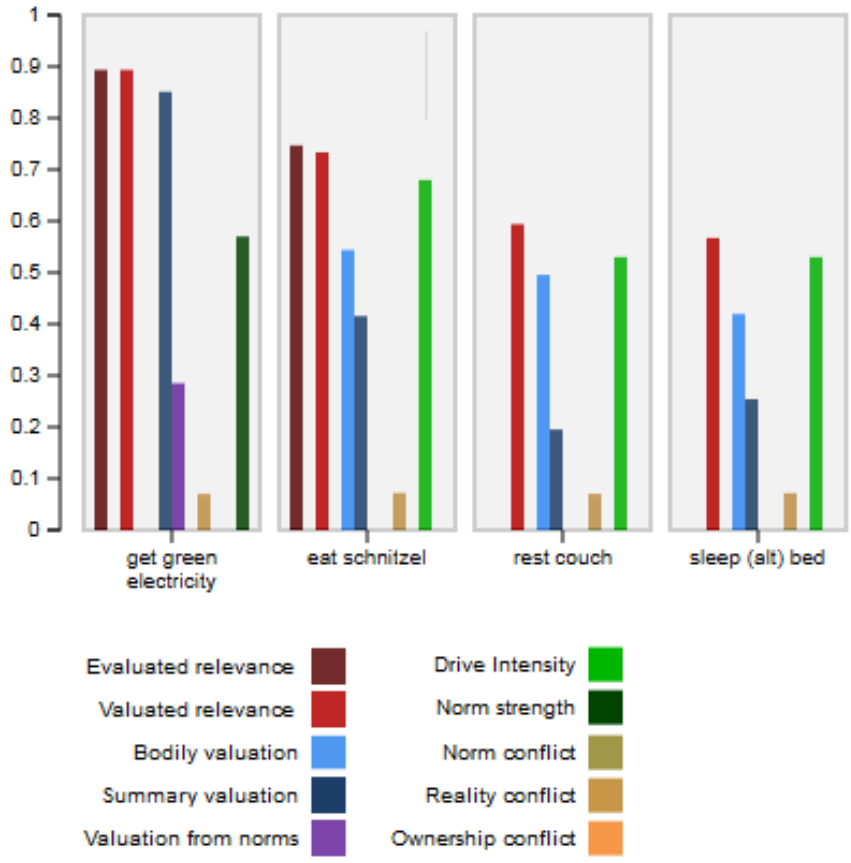

Figure 6.34: Increasing emotional matching enables the impact of confidence in evaluation and thereby leads the agent to get green electricity. Activation of memorized summary valuation of the goal to get green electricity by an compliant current emotional state, activates its (and hence the goals) confidence value. (cf. SC scenario 1.3.1 in Fig. 6.33)

Simulations showed that an increase of confidence is depending on a good emotional matching (i.e., good match between current emotion and memorized summary valuation of green electricity). Confidence is an attribute of the agent's memorized summary valuation (of green electricity). Hence, the better green electricity is activated via emotional activation, i.e., the agent's current emotional state, the higher is the agents confidence about the goal. That is, an increase of memorized confidence without an appropriate emotional matching, would not provide the premises

for the confidence value to work in evaluation. Figure 6.34 shows how an increase of emotional matching together with confidence lead the agent to get green electricity. The dependency on the confidence value is shown by reducing it slightly. In this case another action (e.g. eat due to high drives) would be selected since the evaluation of the goal (but not the valuation) is reduced due to the lower impact of confidence (see Fig 6.35). That is, the memorized summary valuation and its activation by the agent's current emotion has not changed. However, the agent is uncertain about the given valuation, caused by the reduced confidence. That is, he is not sure if the goal would bring him the expected pleasure.

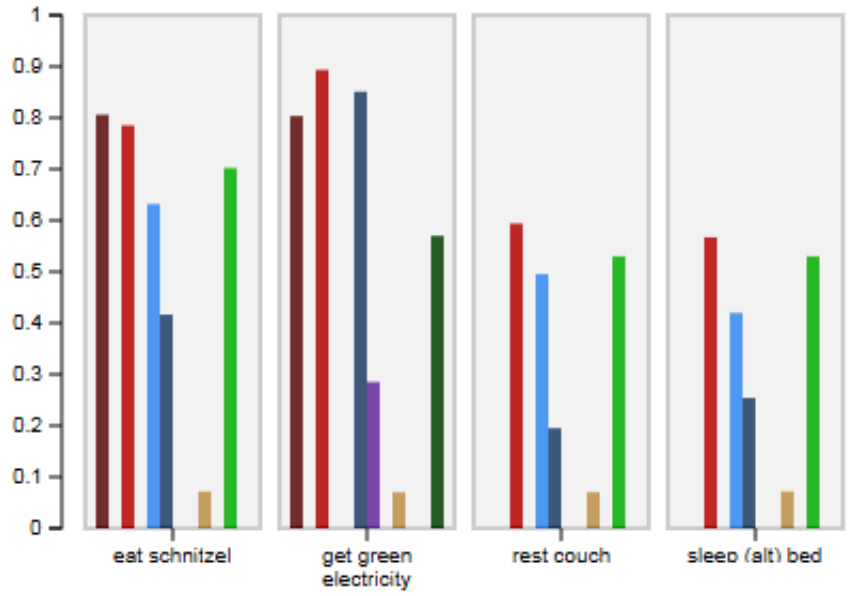

Figure 6.35: Impact on evaluation by reducing confidence. Valuation stays the same, but due to reduced confidence the agent prefers another action (eating) due to the evaluated relevance (cf. SC scenario 1.3.2 in Fig. 6.33).

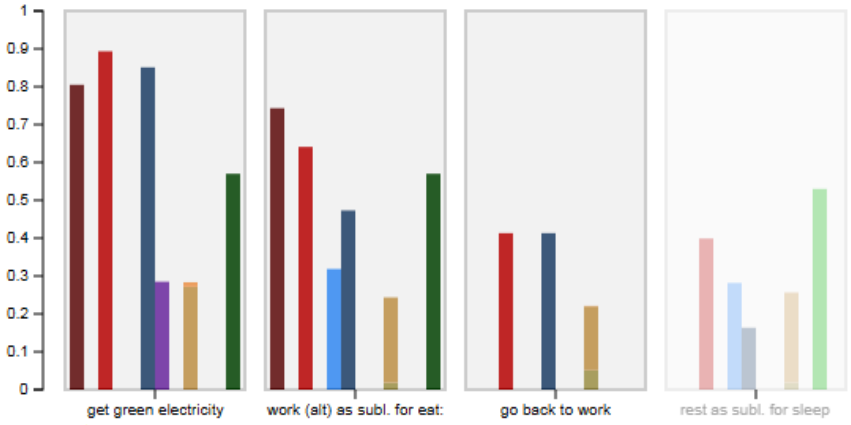

Figure 6.36: A higher conflict sensitivity of the agent is able to compensate the reduction of evaluation due to a low confidence. A higher conflict triggers mediation mechanisms that reduce the valuation of the other - conflicted - goals. Additionally, normative valuations are higher weighted in evaluation. (cf. SC scenario 1.3.4 in Fig. 6.33)

The low impact by confidence can be compensated by causally independent variables, e.g. a higher conflict intensity (see Fig 6.36). The social media message generates a norm conflict, by activating the social norm to act environmentally friendly. Additionally, the recognition of a desired object that belongs to some one else generates a reality conflict (the social media friend that has green electricity and is liked for that by others). Due to the increased conflict sensitivity, mediation mechanisms are activated (e.g. sublimation) towards the goal to get green electricity. Most importantly, the valuation of the other goals are reduced thereby. Additionally, the higher conflict sensitivity increase the relevance for normative valuation in the evaluation mechanism.

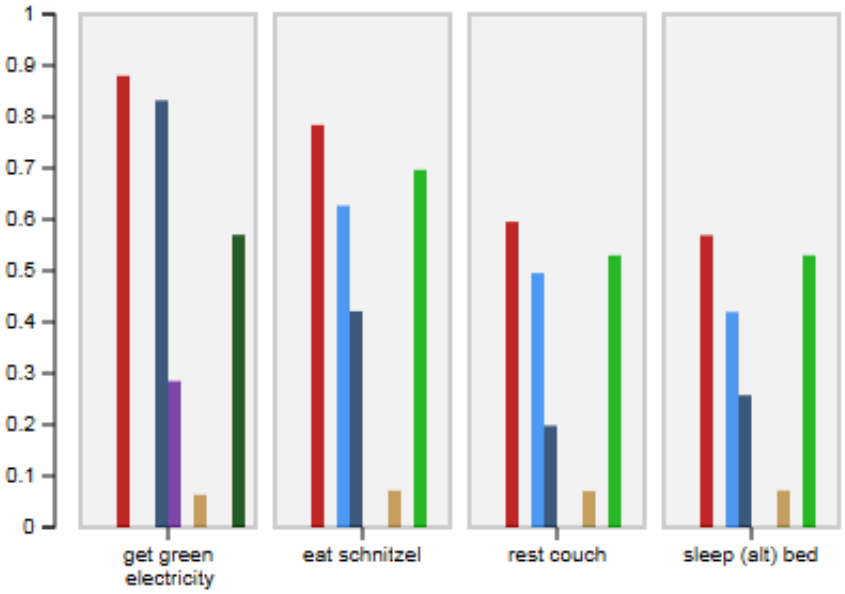

Figure 6.37: By reducing the ego-strenth (via neutralized intensity) the conditions for evaluation are not met any more and the agent decides based on the goals' valuations, where the lower confidence value (Fig. 6.35) does not have an impact any more (cf. SC scenario 1.3.3 in Fig. 6.33)

Generally speaking, the impact of a lower confidence value is only valid because the agent evaluates possible goals. That is, the agent is in a reflective mode of reasoning and hence considers and weights various variables (as described in Chapter 5.6). If the agent is forced to a fast decision or reflection is not necessary or possible (due to low ego-strength or unambiguous goals), the made decision is based solely on the aggregation of a goal's valuations. We show that by reducing ego-strength the 'bad' influence of the confidence value looses its ground, beacause the agent do not evaluate (but only valuate) possible goals. The valuation of the goal, activated by perceiving the message and the agent's current emotion, is high enough and is not relativised by the low confidence value anymore .

However, a higher conflict (sensitivity) is not always beneficial for normative goals. As always, we see that the context (of other model variables) matters and the premises and interdependence of variables and goals must be considered. And as always we cannot give absolute statements about the impact of variables by only considering their variable context, since decision making only sets priorities, i.e., in context of other goals. If we increase the agent's unpleasure (e.g. via drives) the resulting sublimation from conflicted goals to other goals lead to a higher increase of the goal to work. The reason is that the goal to work has a lower reality conflict, since it gets the sublimed valuation of conflicted drive goals (that have a high valuation due to the high drive unpleasure).

That is, since the agent is already working, it is easier for her to continue working and thereby reduce the unpleasure sublimed from conflicted drive goals, instead of the more effortful goal to get green electricity. Moreover, the lower emotional matching does not trigger the goal's memorized summary valuation sufficiently. The change of the agent's current unpleasure implicitly also changes the agent's current state. In result other goals are higher activated than the goal to get green electricity (except we would also change its memorized summary valuation according to the changed emotional state).

The results emphasize the view that - especially regarding evaluation - it is not effective to try to manipulate variables independently, but their relation to each other. This is shown by regarding emotional matching as one variable that considers the relation between memorized valuation and the current state, and by the dependence of the confidence value in these values. In particular, we also see that evaluation could distort the positive effects of valuation regarding environmentally friendly behavior, if not all conditions for evaluation are met. That is, reflective reasoning could distort good valuations of green electricity, by realizing that the good valuations and good fit to the current internal (emotional state) and external (perceiving the message) situation is not certain.

In general, the simulation results demonstrate four means of prioritizing a goal: (1) to increase a direct impact variable, (2) to consider its (indirect) dependence from other variables, (3) to find variables that compensate an negative impact and consider (1) and (2) in this variables, (4) to find variables that disable the negative impact of variables. However, it does not suffice to consider these aspects only regarding the target goal (get green electricity). It has to be considered for all concurring goals.

6.2.6.2 Empathy with the Sending Agent and Nature

We saw the different aspects of emotion and how we can influence decision making via various impact factors. These aspects emphasize the purpose of emotion as an integrative holistic means of decision making, bridging different sources of decision making. Next, we will see that this is also the case regarding empathy as the social aspect of adapting the decision on others bodily expressions. As described in Chapter 4.3.3.8 the descriptive term empathy is used for emotion attribution to other agents and the impact on one owns emotion by that process.

We see in Figure 6.38 that - even if no causal relation to the goal to get green electricity exist - empathy with the other agent may lead the agent to get green electricity. In the demonstrated simulation case scenario this is the case due to low competition with other goals and better matching of the agent's current emotional state with the memorized summary valuation of green electricity. By influencing the agent's current emotion, empathy facilitates appropriate emotional matching. This only happens if the premises of emotional transfer (the mechanism behind 'affective empathy' are set. As shown in SC scnario 1.4.1 these are a strong memorized association to the sending agent (which the agent gets the message from) and a positive summary valuation (including high confidence in valuation) of sending agent. Only then a positive emotional expression in the social media message impacts the emotional state of the receiving agent sufficiently.

We also see the high dependencies between these variables by systematically reducing these single variables (SC scenario 1.4.2), which lead to prioritizing another goal instead of getting green electricity. For instance, if the agent does not know the sending agent (i.e., does not have memories and valuations of him), an important premise-variable is not set for required emotion

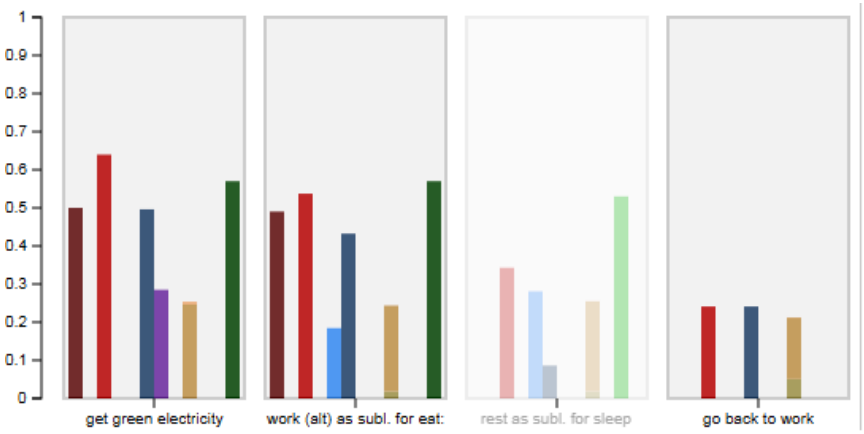

Figure 6.38: Using empathy mechanisms with the sending agent can lead the agent to get green electricity, if competition to other goals is not high and the premise-variables of empathy are high (cf. SC scenario 1.4.1 in Fig. 6.33)

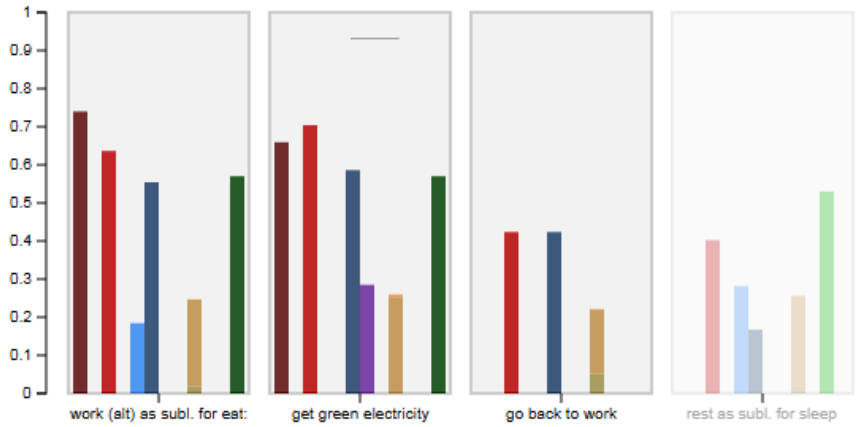

Figure 6.39: The lack of one of empathy's premise-variables suffices to change the agent's decision to not get green electricity, e.g. if the agent does not know the agent that sends him the social media message (cf. SC scenario 1.4.2 in Fig. 6.33)

transfer (see Fig. 6.39). This is also the case for the other premise-variables. To give an example, even if the sending agent is positively valuated, a low memorized association strength to the agent or a low confidence in its memorized valuation suffices to disable the required emotion transfer. This again emphasizes the sensitivity and fragility of decision making, regarding the dependence of impact variables to other variables and to other goals.

However, we see again that a lack of premise-variables can be compensated by increasing one or more other premise-variables of empathy. But compensation can also be reached by another impact factor of the agent's current emotional state, e.g. by perception. In Fig. 6.40 we see how perceiving nature in a bad - i.e., unpleasureful - state is able to compensate the lack of emotion matching by influencing the agent's current emotion.

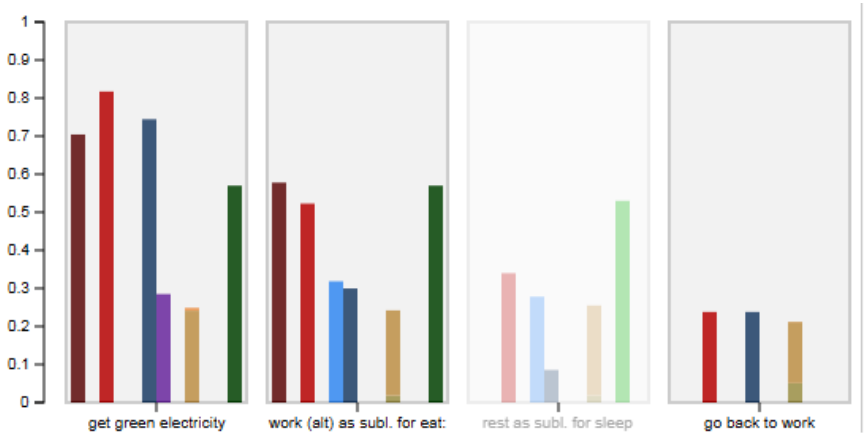

Figure 6.40: Using empathy mechanisms with nature can lead the agent to get green electricity, if competition to other goals is not high and the premise-variables of empathy are high (cf. SC scneario 1.4.2 in Fig. 6.33)

7 Discussion

Many problems and research questions are entailed by selecting and translating theories from psychoanalysis, psychology, social sciences, and neuroscience into a foundational mental architecture for decision making. Such mental architecture considers the foundations of human information processing and bases decision making on it. At the end of this thesis we will see in retrospect how the author was able to deal with these challenges, in particular how the working hypothesis and the according approach was able to serve as a template for fulfilling the thesis' objective. Reaching the objective of developing a foundational mental architecture entails demonstrating its applicability as a flexible and powerful research tool for tackling psychological questions. In case of the thesis at hand, these are questions about basic decision making in social context and questions about environmentally friendly behavior. In particular, we aimed at showing that a translation and validation of theories from the mentioned disciplines is possible. The resulting simulation model is meaningful to support these disciplines in their endeavor to understand human behavior and provide recommendations that serve humans as individuals and as a society.

7.1 Conclusion

How successful was I in reaching the objective of the thesis at hand and how does the result help to reach the overall objective: to reverse engineer the mind for the exploration of socially important questions? Is such an approach possible, necessary, and helpful? In reaching the thesis' objective I aimed at implicitly addressing the following typical questions concerned with this thesis' endeavor.

Why should we develop a computational model of the mind? Can it be a valid representation of the human mind? How can this be helpful? Which value does it add to other approaches and disciplines? These are some of the important questions posed by researchers that are critical of computational models.

Why is a holistic model of the mind necessary? Why is it not possible to abstract from bodily aspects and emotion when information-processing is at stake? Isn't it better to develop a tailored model for a specific research question? These are some of the important questions asked by researchers that are critical of intensive interdisciplinary knowledge translation.

I hope that the detailed answer is given in the thesis at hand, and I will provide a summary in the remainder of this chapter. I will do this by recapitulating (1) how the chosen methodology was able to tackle the challenges in developing the aimed at model, (2) how the model was able

to demonstrate the consistency and validity of the underlying psychoanalytic and neuroscientific theories, and (3) how simulations were able to explore the model and provide insights about the determinants of cooperative and environmentally friendly behavior. In this regard the remainder of this chapter is structured to show how the thesis handled (1) methodological, (2) model, and (3) simulation challenges. Therefore, for each of these areas the following topics are recapitulated.

7.1.1 Methodical Recapitulation

Recapitulating the methodological questions formulated in Fig. 3.14, we can observe how following the methodological criteria and applying case-driven simulation supported to reach this thesis' objective.

The developed methodology is an iterative process with a frequent iteration cycle of analysis, specification, modeling, implementation, software verification, and simulation, and a less frequent iteration with the analysis step (see Fig. 2.4). Technically, case-driven simulation can be regarded as test-driven modeling and implementation. Scientifically, the methodology can be regarded as joining theory-driven and data-driven approaches.

The thesis showed how following a holistic-unified and generative-functional approach enables to develop a explanation model that demonstrates the causal chain from sensory input to executed action. This causal chain considers how decision making is generated from low-level processes. Such an approach not only enables a fine-grained explanation model, but also provides a ground for fine-grained experimentation. This includes observing the dependency between various variables at different levels (e.g. the impact of bodily needs and reasoning about non-body-related goals via neutralized intensity). By that the recognition of unexpected relations between variables and detailed model exploration are facilitated.

A valid representation of the human mind's functioning can only be reached by following an interdisciplinary approach in translating theories from other disciplines into a computational model. Therefore required intensive interdisciplinary collaboration comes with challenges. The developed methodology was successful in identifying requirements posed by other disciplines and translating them into requirements for a computational model. Structured and guided exemplifications of research questions provided an appropriate interdisciplinary platform for communication. It merges a case-study approach (psychoanalysis) with a use-case-driven approach (computer science) and therefore is able to combine methodologies. Simulation cases as a shared specification proved helpful to formulate the processes underlying the aimed at model. Their structure supported conducting interdisciplinary reviews and developing fine-grained test specifications that not only consider black box, but also white box tests. In particular, grouping parameters into determinants and reason about their expected determinability proved as an appropriate means to formulate test-able assumptions from psychoanalysis and other disciplines. By following multiple cycles of case-driven simulations, we were able to sharpen the model incrementally, considering different levels of feedback. Applying the same methodology with different research questions (social interactions, environmentally friendly behavior), enabled condensation of the model into a more functionally unified and compressed version.

The presented methodology helps to determine requirements and solutions at different levels and in a methodical way, which lead to an incremental refinement of requirements and assumptions (i.e., test specifications). Hence, requirements and assumptions of the different steps in the methodology are related to each other. Such requirements state assumptions about the features

that are required to generate the described behavior. Of course, the behavior description itself represent implicit requirements, since the desired behavior is described. But we only use them to analyze the required functions that generate this behavior. Hence, we use the behavioral requirements (e.g. perceiving and focusing on a food source) to deduct – with the help of researchers from other disciplines – the functional requirements (e.g. perception for the fulfillment of desires, attention mechanisms). First the conceptual requirements are analyzed. That is, which concepts are required to generate the described behavior (e.g. motivation based on bodily needs, social norms etc.)? In most cases the process description (of the inner processes) in the exemplary case formulate the required psychological and sociological concepts. But the transformation into a technical model may also trigger requirements. Next, conceptual requirements trigger model design requirements, which the implementable agent model must fulfill (e.g. the interface between body and decision unit, environmental coupling). In developing such model, implementation requirements may feed back on conceptual and/or model requirement (see feedback possibilities in Figure 2.4). This is especially the case, since not only functional requirements, but also data requirements are analyzed. For instance, required parametrization of functions may adapt the simulation cases' defined data determinants (e.g. the incline of hunger).

7.1.2 Model Recapitulation

The developed model was able to address challenges of mental architectures and computational models of emotion (see also Figure 3.14 and Chapter 3.5). The methodology and the usage of two research domains, which tackle the same underlying question of a mental architecture, enable incremental model development. The insights from developing and simulating the conceptual model were used to identify its functional essence. This increased not only the generative-functional aspect of the model, but also the unification of its functions. This results in translating the descriptive concepts of perception, motivation, emotion, feeling, norms, and reasoning into the functional concepts of activation, valuation, mediation, and evaluation. The foundations of human information processing are identified as activation and valuation (mediation and evaluation only operate upon them). Simulations showed that these functions are able to generate and explore human behavior under consideration of the social context. In particular, activation is modeled as the mechanism underlying perception, valuation underlying motivation, mediation underlying social norms, and evaluation underlying reasoning. As a result, the SiMA-C model also has a clear consideration of gradual dual processing, integrating low and high-level processing. Conceptually, this can be regarded as integrating motivation, emotion, and cognition without reducing the model to one focus. This unified and compressed version of the conceptual model (which is based on the SiMA model) also proved easier to handle in exploratory simulation and for empirical mapping.

As expected, using a theoretical abstract framework was helpful in providing a glue for unification. This was especially important due to the usage of assumptions from different theories. Here we can say that the model was able to translate and integrate assumptions from psychoanalysis (Sigmund Freud's metapsychological model), neuroscience (Antonio R. Damasio's theory of emotion and integration of life-regulating mechanisms), neuropsychoanalysis (Mark Solm's theory of affects), and psychology (Simon's, Kahneman's, and Gingerenzer's theories of adaptive decision making using heuristics) and hence show their compliance.

Both, the conceptual model and the SiMA-C model, were successful in considering the analyzed weaknesses in State-of-the-Art (SotA) approaches. In particular, the SiMA-C model considers

foundational decision making under consideration of the social context, which is represented "through the agent's mind" (Cas00) in two ways: (1) Internalized norms as top-down influences in social context. By considering them as (strategic or tactical) goals, they are integrated in the model of activation, valuation, mediation, and evaluation. By considering different activation paths, the conceptual separation of injunctive and descriptive norms is included implicitly in the model. (2) Empathy as bottom-up influences in social context. By considering 'affective empathy' as emotion attribution and emotion transfer, the concept is functionally integrated with a model of emotion. By considering 'cognitive empathy' as an extension of emotion attribution under consideration of action recognition and planning, different levels of empathy are integrated. This also emphasizes how a computational model is able to bridge psychological and sociological theories. Another missing topic in SotA models is the lack of following an evolutionary approach in using emotion as a base for decision making. Such approach would show how different layers of life regulating mechanisms could be integrated and unified. Hence, different than conventional cognitive architectures, the SiMA-C model was successful in identifying the information-processing-mechanisms in psychological models and using computational means to represent them. This includes showing how an foundational unified model of life-regulating mechanisms is able to explain heuristic processing in adaptive decision making. Different to many approaches to emotion, the SiMA approach does not use folk psychology as a starting point to define emotion. The chosen generative-functional approach show how emotion extend the means for activation and valuation for better adaption to the dynamic environment. In this regard, emotion as a holistic state indicator is able to mediate between internal and external world, and between different internal demands. Such integrated state indicator not only represents the agent's current state, but also its expected change. The indication of one's state to other agents enables additional means for adaption. Opposed to wide-spread computational models of emotion, they are not modeled as consequence of cognition, but rather as a means of cognition. That is, emotion is used to model cognitive processes, not vice versa. This is the case both for low-level cognition, where emotion is used for memory activation, and for high-level processing where they are used as feelings for reflective thinking - applying the expectations from memories on the current internal and external context. In summary, using the functional concept of a state indicator enables to integrate the descriptive concepts of drives, emotion, feeling, and their usage in activation, valuation, mediation, and evaluation for low and high-level mental processing. Different principles (valuation and evaluation) and scopes (local in valuation and global in evaluation) in using the state indicator are able to account for various facets in human information processing.

The SiMA-C model also shows the hierarchical relation between the unconscious and the conscious, motivation and emotion, emotion and feelings. All these concepts are data that are processed differently. That is, these dualistic dialectics are data-based based, not function-based. For instance, the SiMA-C model shows how emotion as a state indicator provides holistic and integrated information how to enhance our state. But how we use this information may differ. If the internal and external situation allows for it[1], the most efficient usage of the state indicator is as exactly that: a concept that *indicates* the need of further mental processing - it may indicate[2] its direction, but not the details. The state indicator should be regarded as a means for the purpose of conscious reflection, not *as* a final decision. In particular, it should not be used as a final decision making basis, but it has to be related to all aspects of the current situation, which may differ from the strict memory-based valuation (e.g. long-term considerations or a different

[1] Cognitive resources or external constraints (e.g. the need of fast decisions) may hinder this process.

[2] See the etymology of 'indicate' as 'to point out', with 'dicere' meaning to say something. In this regard, emotion as a state indicator 'tells us' something about our state - we only have to listen (consciously).

current agent's state). Even when the state indicator is conceptualized as providing information for decision making, it does not determine decisions. Not the state indicator, but its usage as drives, emotion, and feelings determine the decision[3]. Hence, the drive-based valuation of goals can be changed by emotion-based valuation, and both valuations can be changed by feelings[4]. But these changes are only extensions. Higher layers of valuation do not invalidate the valuations of lower layers, but only put them in another light (by considering additional information, e.g., unpleasure, norms, the emotional state of others etc.). The SiMA-C model proposes a non-proportional aggregation of valuations in the primary process, and a focused aggregation in the secondary process. However, the details of the relation between different valuations can still be regarded as an open research question. The SiMA-C model not only supported the demonstration of the relation between drives, emotion, and feelings as means for valuations; the need for consistent modeling in a holistic frame enabled their integration in the first place. Furthermore, simulations of the SiMA-C model enable the observation of the internal and external dynamics of their interactions. We are able to observe how drives, emotion, and feelings evolve and operate during simulation; how they dependent on each other and how they determine the agent's decision. These possibilities of simulations with the SiMA-C model demonstrate its contribution in relation to empirical research.

7.1.3 Simulation Recapitulation

The simulation results show the usefulness of the SiMA-C model's possibilities of parametrization and causal comprehension. The simulation case approach enables model parametrization in a well-structured manner. The simulation case specification not only help to break down the complexity in incremental development and evaluation, but more importantly, it facilitates the formulation of psychological hypotheses into a simulatable test specification. This fine-grained testability of hypotheses is arguably only possible in computational models. Simulation results demonstrate how the SiMA-C model enables experimentation with internal variables that cannot be manipulated directly in empirical experiments. Especially the ability to track how a parameter change lead to behavior change, show how we are able to test strong hypothesis about variables' impact. In the end, the thesis at hand showed how these hypotheses (that are incorporated in each simulation case scenario) are consistent with the underlying theories. Most importantly, the hypotheses do not claim to be the only valid explanations, but only candidate explanations. This is emphasized by showing how different variable changes may lead to the same behavior: The same behavior is determined differently in (1) different external situations (2) by the same or different personalities (3) with the same or different internal state. That is, changes in one or more of these three groups may lead to the same behavior (e.g. the agent may share the Schnitzel because of anger or because of joy). That is, the simulation case specifications aimed - amongst others - at showing that the same personality can be determined to execute action x because of different reasons (e.g. due to different internal states), and that the same environment determines the same action due to interaction with different internal variables, and that the same internal variables may lead to different behavior in different environments, etc.

The possibilities demonstrated by the simulation cases emphasize that decisions are not determined mono-causally. Even if a change of only one determinant suffices to change the agent's

[3] In general, not information per se, but how we use it, is crucial.

[4] When considering the dependence between the agent's state and goal-valuations, SiMA-C is able to extend Hume's metaphor of 'reason as a slave of emotion': Reason (or rather decisions) may be originally a slave of emotion, but it is able to buy itself freedom into lighter forms of dependencies (e.g., employer or freelancer).

behavior, only the determinant's dynamic interaction with other variables determine the agent's behavior. This may be obvious due to the causal relations between model variables (e.g. drives and emotion), but also due to indirect causality (e.g. drives and reflective thinking with feelings via neutralized intensity) , which is difficult to analyze prior to simulation. Simulations of a mental architecture with such parametrization-possibilities are arguably better able to demonstrate the dynamics of decision making than pure analytical approaches. This includes the dynamic between model variables and between model and environmental variables. However, for future work, to harness the potential of the simulation case approach, it should be extended with extensive data analysis (data mining etc.). However, due to the high number of variables, at the beginning a guiding experimentation using simulation cases is helpful. After the model has established itself in controlled experiments, an analysis may show which parts of the model need further intensive data analysis.

Simulations also help to emphasize the central role of some determinants. This is the case for memorized valuations and neutralized intensity. However, simulations also show that their determinability is dependent on the variable context. Simulations showed how the SiMA-C model can be used to identify and corroborate pre-requisites of impact factors. For instance, simulations show how confidence in valuations (e.g. of green electricity) can be dependent on emotion matching if the perceptual activation is low.

The dependencies of model variables is also shown by the different usage of emotion in the model. The same emotional state can determine different, even 'opposed' behavior, when used differently in valuation and evaluation. That is, an angry agent choose the beating goal when in a reactive mode. In case of sufficient neutralized intensity in combination with sufficient super-ego strength, the same anger can lead the agent to share due to anger that is defended into guilt. If super-ego strength does not suffice, nevertheless the same anger may lead to sharing, if neutralized intensity is high enough to enable the agent to reflect on the situation, using emotion as feelings. That is, emotion may lead to unsocial behavior, if their premises for social behavior are not met. However, they are required for social behavior, but only used in the 'right' way, i.e. they facilitate social norms and reflective thinking.

Simulations also demonstrate how emotion as a holistic state indicator is able to integrate these dependencies. We saw the different aspects of emotion and how we can influence decision making via their various impact factors. This emphasizes the purpose of emotion as an integrative holistic means of decision making, combining different sources of decision making. We see that this is also the case regarding the social aspect of adapting to others bodily expressions. Considering parametrization, in future simulation case specifications we should use emotion as an intermediate-step, i.e. specifying (1) the specific emotion that is expected to determine behavior (and why, e.g. due to which activation paths and valuation principles), and (2) how this emotion is generated, i.e., which determinants cause this emotion.

Such a holistic approach emphasizes that determinants and their values only have relative (not absolute) impact on decisions. That is, simulation demonstrated that not the absolute values of determinants are relevant to specify, but the relation to their dependent determinants. This is also the case regarding goals: Not the expected absolute relevance of a goal determines the agent's behavior, but only the relation to other currently possible goals.

On the one hand, the various impact factors of emotion may increase the fragility of decision making. On the other hand, additional factors increase possibilities of adaptation, e.g. by increasing activation pattern for memories, which enable a fine-grained match with the features of the current internal and external situation. Of course additional impact factor - as with any

impact factor - require that a policy need to take care of it. For instance, considering emotion and intention attribution provide additional means to reach the target behavior. But it also may hinder that. For this specific example, simulations show how the same body expression may lead the agent to beat or share, depending on the memories about such body expression and the valuation of the other agent (which is based on memories about the other agent and the agent's current internal state). That is, the mechanisms behind 'empathy' may lead to social or aggressive behavior depending on the determinant constellation.

The degree of dependencies between determinants and variables, which we saw in simulation, require further control in specifying simulation cases. Therefore a dependency graph that explicate implicit causalities between model variables and demonstrates the branched causal chain of variables could be beneficial. Such dependency graph would also support the specification of simulation experiments.

We also see the high dependencies between emotion's impact factors in systematically reducing these single factors. This is also the case for the other premise-variables. For instance, even if the sending agent is positively valuated, a low memorized association strength to the agent or a low confidence in its memorized valuation suffices to disable the required emotion transfer. This again emphasizes the sensitivity and fragility of decision making, regarding the dependence of impact variables to other variables and to other goals.

Simulations showed how the target behavior (e.g. sharing) is determinable in different ways. This can serve as informing policies. In simulations the easiest way to get the target behavior was to change an agent's memories, since the personality is the only factor (out of the three determinant groups) that does not change during simulation. This is obviously not the case for the perception of the external environment and the agent's internal state, which changes with every simulation step. Hence, in simulation the personality is the best controlled determinant group. However, in reality this is not the case. Thus we can only use these insights from simulation to know how to adapt the external environment and (indirecly) the agent's state on a specific personality, which in the SiMA approach is a fluent concept specified by all valuated memories and a few personality parameters. However, we saw how - for a specific personality - the external state or the internal state can be changed to motivate people to act pro-socially and environmentally friendly. By aiming at changing the internal state (via external perception or enabling drive satisfaction), we can indirectly determine which aspect of a personality will be in the fore, i.e., which memories will be activated (anger activate other memories than joy - in the same personality). This shows, again, that the 'dialectic' between internal state and memories is more efficient than aiming at changing memories. Many simulation case scenarios showed how such priming of an agent is able to determine behavior change via changing the emotional state (e.g. by changing the body expressions of other agents).

Another insight that is able to inform policies that aim to motivate cooperative and environmentally friendly behavior is information about stable and fragile determinant constellations. Many simulation case scenarios showed how in specific determinant constellations a small change of a determinant is able to change the agent's behavior. We saw that in switching from beating to leaving, beating to sharing, and eating to sharing, but also in scenarios for deciding on green electricity. The used model is able to provide an explanation for this behavior dynamic and its fragility. The model is also able to account for and explain counter-intuitive behavior, such as why an increase of anger would lead an agent to leave or share instead of eating the Schnitzel alone or even beating the other. In particular, it is able to explore the premises for such behavior (high neutralized intensity or medium neutralized intensity and high super-ego strength).

The simulations show how, in every case, policies have to address valuation mechanism. That is, policies should emphasize the relevance of behavior for people and society. It would not suffice to do that on a theoretical basis by facts about a topic (e.g. green electricity), but also by relating it to the agent's valuation sources (norms and (bodily) demands)[5]. The according valuation of a topic or goal can be addressed by choosing the 'right' activation path (perception, norm, body).

A general statement regarding variable manipulation via determinant was the result of a global view on the simulation cases: In general the simulation results emphasize that to prioritize a goal we can (1) increase a direct impact variable, (2) consider its (indirect) dependence on other variables and try to change them, (3) find variables that compensate an negative impact and consider (1) and (2) in these variables, (4) find variables that disable the negative impact of variables. But it does not suffice to consider these aspects regarding the target goal (sharing or getting green electricity), but also for all concurring goals. In particular the target behavior can be motivated not only by increasing the determinants of the target goal, but by reducing the determinants of concurring goals.

7.2 Outlook

The thesis at hand aimed at showing that the used interdisciplinary approach is able to provide a fine-grained holistic model of human information processing. The results are helpful for developing a tool to address socially relevant questions. At the end of this thesis we take an outlook regarding further model development and applications.

7.2.1 Future Model Development

The developed model show possibilites for further development. Some of them are mentioned next to sketch where the way may lead.

Towards Further Unification and Distributed Control

The incremental unification and simplification of a functional model proved beneficial for implementation, validation, and exploration. By starting from a full-blown model we were able to analyze and learn how the same functionality can be reached by a simpler, more manageable model. This process of model sharpening will be continued and first ideas are already in the process of examination[6]. With the aim to simplify activation and valuation further, both are compressed into a model of activation (see Fig. 5.3). The generic model considers activation by a source and derivation of that activation by a manipulator. Conceptually, the source can be regarded as a demand or affordance, and the manipulator can be regarded as a norm posing conditions on how to fulfill the demand. The impact of the activation source and manipulator is determined by their current strength and the activation and manipulator rule, respectively. The current strength can be regarded as the source's and manipulator's current relevance. In case of a demand, this is the underlying lack. In case of a norm, this is its activation by a conflict.

This concept is applied in a distributed approach for a human-inspired control system for energy-regulation in buildings. Due to the number of different physical components (and their interactions) and the different structures of buildings, a distributed approach is assumed to be helpful

[5] In this regard, we can state that emotion is able to operationalize knowledge.

[6] The following concept was developed by this thesis' author for a human-inspired control system for energy-regulation in buildings.

in controlling the system's complexity. For that the agent paradigm is used for every component of the control system (including the controlled components). In this regard, low-level primary process (PP) agent and high-level secondary process agents are distinguished. Out of the local control and interaction between PP agents, global control is expected to emerge. Only when necessary and possible, SP agents would further use and extend the result of PP agent control. Hence, the control cycles of PP agents and SP agents are de-synchronized. First results with focus on control by PP agents showed the feasibility of this approach (KSSW16). Agent-based composition of the system proved helpful regarding the extensibility and flexibility of the system. PP-agent-based control, which was easy to implement using the above sketch activation concept, proved sufficient for standard situations (i.e., every-day control). We will further test the limitations of PP-agent-based control, i.e., when SP-agent-based control is required.

Towards a Model of Attitude

The developed SiMA-C model is used to explore foundational human information processing. Aspects where a computational approach is especially beneficial include (1) showing how decisions are based on 'unconscious' data (based on memories' associativity), (2) grounding norms in demands, (3) implementing mechanisms of heuristic decision making, and (4) demonstrating how the dynamics in decision making generate dynamics in society through frequent agent interactions in large-scale social simulations[7].

An important question is how to use and extend the SiMA-C model of activation, valuation, mediation, and evaluation to represent attitude and its usage in decision making. This is especially relevant in social context. Decisions often can be regarded as attitude-based. Two aspects are particularly interesting. (1) How attitude is expressed differently dependent on the current (internal and external) situation, and (2) how such an adaptive attitude is influenced and formed by social interactions. Identifying the determinants and mechanisms of attitude is expected to support formulation of policies for environmentally friendly and pro-social behavior (see future applications below).

Towards Feasible Data Storage and Integrated Software Tests

A crucial aspect of the SiMA-C model is its data structures. The creation and configuration of the underlying data bases was a hurdle for simulation. This is particularly the case when aiming for automatized simulation experiments and extended data analysis. An analysis of data base frameworks led to choosing a conventional relational data base (RDB). The main reasons were the availability of established frameworks and object relational mapping (ORM), which is able to mask and abstract from the details of how the data is stored, and enable the usage of java classes. However, experience with the (interdisciplinary) requirement for high associative data showed that an RDB with ORM is not able to sufficiently account for the conceptual requirements (stemming from psychological theories) on data structure. The development in object-related data bases and triple stores may support tackling these challenges, since their scalability is increased by now and their frameworks are established. The high associativity also posed a challenge in identifying errors in software tests and model evaluation (cf. the levels of feedback in Fig 2.4). Highly associative data structures and their processing in a mental architecture is expected to support integrating the associative aspect of the brain with the symbolic aspect of the mind.

The different feedback levels in model evaluation also showed the importance of software testing. This enhances identifying the level of feed back, i.e., why the specified expectations did not hold

[7]For the purpose of computational feasibility for such simulations, the model has to be simplified (see concept above).

in simulation. Interweaving software and model testing and incremental evaluation is expected to support this process. In particular the simulation cases should be used as point of departure for detailed software tests. Only after successful software test, the next level of evaluation, model validation would be conducted. In many cases, successful software testing would imply successful model validation.

Towards Further Exploration of Environmentally Friendly Behavior

The thesis at hand demonstrated the suitability of its approach and examined its potential and usage to explore impact factors for cooperative and environmentally friendly behavior. After showing the validity of our model in simulations and empirical calibrations, we will use the simulation to gain further insights about the impact factors for switching to green electricity. For that, we will use the identified different agent types and in a first step conduct a sensitivity analysis to identify the impact of model parameters and variables. This will inform how to formulate a social media message, adapted on the respective agent type. The impact of such a message on the decision to switch to green electricity will be tested in simulation experiments. Using multi-agent simulations we will analyze how the spreading of green electricity is influenced by a marketing strategy with the proposed adapted messages. In a next step, the resulting message will be tested in empirical studies. An expected outcome will enable adopting messages to the user by identifying what type of green electricity consumer she is, using her social media profile. Such a follow-up project requires further development of the model and the simulation tools. Overall, the thesis at hand confirmed the main contribution of simulation for the given research question: Using simulations to explore and examine the impact of variables that are difficult to manipulate in empirical studies, e.g. the external environment, or are not accessible at all, e.g. unconscious impact on decision making. Additionally, simulation – after empirical calibration – informs and guides further empirical research and theory development. In this regard, we will use simulation experiments to test the impact of descriptive norms in different contexts, which will test the found minor impact of social pressure in the used survey and possibly inform further empirical elaboration on the topic. These experiments in turn will be extended by natural pressure as a new variable, which is difficult to manipulate in empirical studies.

7.2.2 Future Applications

Using a mental architecture in simulations enables the recognition of impact factors for specific domains (e.g. cooperative and environmentally friendly behavior). In future work, these insights can be used to develop policies that address these variables to motivate people to act pro-socially and in an environmentally friendly way.

A general objective in this regard is to use the means of simulations to get knowledge how to increase awareness about these topics efficiently and anchor pro-social and environmentally friendly behavior in psychologically relevant variables. In the end, we have to test how well such simulation-generated policies work in real life to enhance social solidarity and environmental awareness.

The research profile behind this endeavor can be described as 'human-centered policy development and application in simulations and (virtual) reality for the promotion of empathy and norms'. As described, simulations with mental architectures are able to inform policies, which can be implemented in different ways. One way is serious gaming: Identifying with one's avatar and recognizing the consequences of behavior can address and support promoting empathy and

norms. The next level of identifying with other people and nature is participating and undergoing according situations in virtual reality. Topics where such methods seem helpful is the inclusion of refugees, and the socio-technical integration of users and e-vehicles.

Realistic agent-based simulations for the examination of policies to support the inclusion of refugees

Europe is confronted with multi-faceted challenges in dealing with the migration of refugees. Developing policies that address psychological mechanisms - such as empathy and shared norms - to support acceptance and readiness for inclusion, support cooperative behavior between current and future new citizens, and hence for successful co-living and co-working.

We will use a prototype agent-based model to examine the impact of gossip, media coverage, and personal experience on the agents' attitude. We will base the prototype model on established psychological assumptions, focusing on mechanisms of pro-social behavior and norms. Social simulations with the prototype model and empirical experiments will inform which impact factors and mechanisms are required for specifying the model using the SiMA-C mental architecture.

Using this framework, empathy and norms, as well as their interaction, will be operationalized and become predictable with a model of activation and valuation. Such a psychologically plausible model enables developing policies. In particular it enables testing the policies applicability to adapt the attitude of both, current citizens and refugees. This would nurture the abilities of acceptance of inclusion on the one hand and the readiness for inclusion on the other hand. Possible policies aim at relating refugees to one self and society in existing experiences and norms, and thus support empathy (i.e., perspective taking) and norm adaption. The resulting policies will be further tested in empirical control studies.

For a first approach to the topic we will use the SiMA-C model to examine a simplified instance of opinion dynamics (without learning). This helps us to analyze the appropriateness of our approach and will provide adaptation requirements of applying the SiMA-C model to this topic. For the inclusion of refugees the attitude-based opinions, people have about each other, are crucial. In the last months we have seen a switch in public opinion about refugees. Media coverage and gossip can be considered as one cause. A research question in this regard is if a model of emotion and norms (as in the SiMA-C model) is able to represent the causal chain in attitude-based decision-making and explain the dynamics of opinion influencing, in particular the opinion change about accepting refugees? Which counter measures may work (e.g., would it help to address mechanisms of empathy, fear, and norms)? The working hypothesis is that opinion dynamics, i.e., the change of opinions, is not unconditionally caused by a change of attitudes, but can be explained by a change in perceiving the current situation (via (social) media). That is, it is unlikely that people's attitudes have changed significantly in the last months, but it is based on how they perceive the topic.

User Modeling

With the premise that the electricity grid will be overloaded due to the high number of e-vehicles, one question is how to get user acceptance for interrupted charging of one's e-vehicl. One possibility could be to apply charging strategies that are adapted to the user and the current situation. Therefore simulations of a user model based on the SiMA-C model can be helpful. Concrete questions that should be answered in user simulations could be: (1) How are charging strategies able to adapt to the user's personality, demands, goals, and norms. (2) How can the charging

application and its user interface help to involve the user emotionally and build trust for semi-automatic charging? In general, simulations of user interactions to predict user acceptance is an application field, where a holistic and functional model of emotion and decision making is helpful.

In order to adapt the existing SiMA model to the problem, a requirement analysis of the user-relevant mental functions will be done. This includes identification of user types and their parameterization. After analyzing these requirements, a user survey will be performed. Furthermore, a study will be conducted to identify possible incentives and decision structures for the identified user types. As a result, various nudging strategies will be tested empirically and the empirical data will be mapped into the SiMA model.

7.2.3 Towards More Interdisciplinary Collaboration

The thesis at hand aimed at showing that the combined interdisciplinary methodology is necessary and beneficial for the aim to develop an artificial mind. We hope that colleagues from the involved disciplines are convinced of this thesis' contribution for their field. The resulting tool can be regarded as supporting psychological and neuroscientific research for theory development and evaluation, but also for providing a tool to operationalize their theories for the sake of socially relevant questions. Simulation of mental architectures should only be considered as an extension of these disciplines' methods and tools. Of course, using an information-theoretic approach is able to provide feedback about a theory's consistency, but cannot replace neuroscientific and psychological (including psychoanalytical) models of the human information-processor. As a result, theories of the mind would benefit from the possibility of a generative-functional and holistic exploration of their assumptions. In this regard, the SiMA approach can serve as a framework to integrate different theories into a holistic model of the mind. In particular, using a layered approach the SiMA framework can be used to define the interfaces between neuroscientific models and psychological models (see Chapter 3.2.6). A layered approach also emphasizes the analytical relevance of considering different perspectives on our research topic by different disciplines, with only different focus. By providing a tool to define the interface between these disciplines and a detailed test of the result by tracking the causal chain from sensor input to action, an important milestone in understanding the human information processor would be reached.

To reach the objective of advancing the methodologies and results of the involved disciplines, further integration of the disciplines is necessary. This thesis hopes to have demonstrated a possibility how to achieve further integration, but on a long-term basis the presented methodology has to be further intensified. In recapitulation one can observe that the development of a structured interdisciplinary methodology was the bottleneck of the thesis at hand. In summary, by following a strict interdisciplinary approach using the SiMA approach, the scientific nature of theories from neuroscience, psychology, and the social sciences is enforced by providing means of model specification and evaluation that goes beyond the limits of these disciplines, which is caused by the constrained accessibility and subjectivity of the mind - constraints that do not exist in an artificial mind. This not only concerns the possibility to manipulate otherwise restricted variables of the mind, but also the wide possibility of predictions, given by the high numbers of parameters and variables. For instance, unlike to having the need of categorizing personalities, simulation of the SiMA model enables varying individual personality-dependent parameters and hence explaining and predicting a wide range of personality-dependent forms of behavior.

The benefits of the chosen approach are already sketched in the thesis at hand (e.g. see Chapter 1.2.1) and have low limits. As described above, the next step is to use the SiMA-C model to

explore how to increase awareness about pro-social and environmentally friendly topics efficiently and anchor pro-social and environmentally friendly behavior in psychologically relevant variables using serious gaming and virtual reality. In the end we have to test how well such simulation-generated policies work in real life to enhance social solidarity and environmental awareness. A first step will be to provide recommendations based on simulations and implement them in serious gaming and virtual reality to anchor them in target groups (e.g., to enhance empathy and norms, and to learn environmental and pro-social patterns of thinking and behavior), with the overall objective to serve humans as individuals and as a society. However, the increasing opportunities of detailed analysis, evaluation, and implementations of means to convince and drive people to a specific behavior that is argued to be beneficial for human individuals and society, comes with high potential of manipulation (e.g. to increase profits). But this risk cannot and should not be the consideration of involved (and hence biased) researcher. The assessment if and how a scientific endeavor serves society has to be conducted with other criteria than scientists use for their objective, i.e., ethical and non-scientific criteria. This can only be done by independent commissions (cf. ethics commissions in biotechnology) that have to include the valuable consulting of scientists and philosophers.

Literature

[ABB+04] Anderson, J. R. ; Bothell, D. ; Byrne, M. D. ; Douglass, S. ; Lebiere, C. ; Qin, Y.: An integrated theory of the mind. In: *Psychological Review* 111 (2004), Nr. 4, S. 1036–1060

[AFR05] Amanta, R. S. ; Freed, A. R. ; Ritter, F. E.: Specifying ACT-R models of user interaction with a GOMS language. In: *Cognitive Systems Research* 6 (2005), Nr. 1, S. 71–88

[ALD+05] Aylett, R. S. ; Louchart, S. ; Dias, J ; Paiva, A. ; Vala, M.: FearNot! - an experiment in emergent narrative. In: *International Workshop on Intelligent Virtual Agents*, 2005, S. 305–316

[And07] Anderson, J. R.: *How Can the Human Mind Occur in the Physical Universe.* Oxford University Press, New York, 2007

[Ari00] Ariely, D.: Controlling the Information Flow: Effects on Consumers Decision Making and Preferences. In: *Journal of Consumer Research* 27 (2000), Nr. 2, S. 233–248

[Ari09] Ariely, D.: *Predictably Irrational, Revised: The Hidden Forces That Shape Our Decisions.* Harper, New York, 2009

[AT06] Axelrod, R. ; Tesfatsion, L.: Guide for Newcomers to Agent-Based Modeling in the Social Sciences. In: *Handbook of computational economics 2.* North-Holland, Amsterdam, 2006, S. 1647–1659

[Axe03] Axelrod, R.: Advancing the Art of Simulation in the Social Sciences. In: *Japanese Journal for Management Information Systems* 12 (2003), December, Nr. 3, S. 1–19

[Axe06] Kap. AGENT-BASED MODELING AS A BRIDGE BETWEEN DISCIPLINES In: Axelrod, R.: *Handbook of Computational Economics.* Bd. 2. Elsevier, 2006, S. 1566–1584

[BA08] Becker-Asano, C.: *Wasabi: Affective Simulation for Agents with Believable Interactivity*, University of Bielefeld, Diss., 2008

[Bac09] Bach, J.: *Principles of Synthetic Intelligence: PSI: An Architecture of Motivated Cognition.* Oxford Univ Press, New York, 2009

[Bac11] Bach, J.: A Motivational System for Cognitive AI. In: Schmidhuber, J. (Hrsg.) ; Thorisson, K. R. (Hrsg.) ; Looks, M. (Hrsg.): *Proceedings of the 4th International Conference on Artificial General Intelligence*, Springer, Berlin, 2011

[Bar06] Barret, L. F.: Are Emotions Natural Kinds. In: *Perspectives on Psychological Science* 1 (2006), S. 28–58

[Bat94] Bates, J.: The role of emotion in believable agents. In: *Communications of the ACM* 37 (1994), Nr. 7, S. 122–125

[BC95] Baron-Cohen, S.: *Mindblindness: An Essay on Autism and Theory of Mind.* The

MIT Press, Cambridge, MA, 1995

[BC99] Bargh, J.A. ; Chartrand, T.L.: The unbearable automaticity of being. In: *American psychologist* 54 (1999), Nr. 7, S. 462–479

[BC12] Baron-Cohen, S.: *Zero Degrees of Empathy.* Pinguin Books, London, 2012

[BD98] Bartl, C. ; Drner, D.: Comparing the Behaviour of PSI with Human Behaviour in the BioLab Game / University of Bamberg: Lehrstuhl Psychologie II. 1998. – Memorandum Nr. 32

[BD05] Bechara, A. ; Damasio, A. R.: The Somatic Marker Hypothesis: A Neural Theory of Economic Decision-Making. In: *Games and Economic Behavior* 52 (2005), S. 336–372

[BDZ+13] Bruckner, D. ; Dietrich, D. ; Zeilinger, H. ; Kowarik, D. ; Palensky, P. ; Doblhammer, K. ; Deutsch, T. ; Fodor, G.: ARS: Eine technische Anwendung von psychoanalytischen Grundprinzipien fuer die Robotik und Automatisierungstechnik. In: *Psychoanalyse im Widerspruch* 50 (2013), S. 57–116

[Ber04] Bersini, H.: Whatever emerges should be intrinsically useful. In: *Artificial Life* 9 (2004), S. 226–231

[Ber14] Bermdez, J. L.: *Cognitive Science. An Introduction to the Science of the Mind.* Cambridge University Press, UK, 2014

[BF72] Block, N. J. ; Fodor, J. A.: What Psychological States are Not. In: *The Philosophical Review* 81 (1972), Nr. 2, S. 159–181

[BGSW13] Bruckner, D. ; Gelbard, F. ; Schaat, S. ; Wendt, A.: Validation of cognitive architectures by use cases. In: *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE)*, 2013, S. 1–6

[BKVH08] Bruckner, D. ; Kasbi, J. ; Velik, R. ; Herzner, W.: High-level hierarchical semantic processing framework for smart sensor networks. In: *roceedings of IEEE HSI*, 2008

[BN71] Bell, C. G. ; Newell, A.: *Computer Structure: Readings and Examples.* McGraw-Hill, New York, 1971

[Bon02] Bonabeau, E.: Agent-based modeling: Methods and techniques for simulating human systems. In: *PNAS* 99 (2002), S. 7280–7287

[Bra84] Braitenberg, V.: *Vehicles: Experiments in synthetic psychology.* MIT Press, Cambridge, 1984

[BT11] Borrill, P. ; Tesfatsion, L.: Agent-Based Modeling: The Right Mathematics for the Social Sciences? In: Hands, J. B. Davis & D. W. (Hrsg.): *Elgar Companion to Recent Economic Methodology.* Edward Elgar Publishers, New York, 2011, Kapitel 11

[BVCP13] Bravo, G. ; Vallino, E. ; Cerutti, A.K. ; Pairotti, M.B.: Alternative scenarios of green con-sumption in Italy: An empirically grounded model. In: *Environmental Modeling Software* 47 (2013), S. 225–234

[Can98] Canamero, D.: Issues in the design of emotional agents. In: *Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium*, 1998, S. 49–54

[Cas98a] Castelfranchi, C.: Emergence and Cognition: Towards a Synthetic Paradigm in AI and Cognitive Science. In: Coelho, H. (Hrsg.): *IBERAMIA, LNAI 1484*, Springer-Verlag, Berlin Heidelberg, 1998, S. 13–26

[Cas98b] Castelfranchi, C.: Simulating with Cognitive Agents: The Importance of Cognitive Emergence. In: *Multi-Agent Systems and Agent-Based Simulation.* Springer-Verlag, Paris, 1998

[Cas00] Castelfranchi, C.: Through the agents' minds: Cognitive mediators of social action. In: *Mind & Society* 1 (2000), Nr. 1, S. 109–140

[Cas06] Castelfranchi, C.: Cognitive architecture and contents for social structures and interactions. In: *Cognition and Multi-Agent Interaction*. Cambridge University Press, New York, 2006

[CG04] Cialdini, R. B. ; Goldstein, N.J.: Social Influence: Compliance and Conformity. In: *Anual Review of Psychology* 55 (2004), S. 591–621

[CGG01] Clore, G. L. ; Gasper, K. ; Garvin, E.: Affect as information. In: Forgas, Joseph (Hrsg.): *Handbook of affect and social cognition*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2001, S. 121–144

[Coo07] Cooper, R. P.: The Role of Falsification in the Development of Cognitive Architectures: Insights from a Lakatosian Analysis. In: *Cognitive Science* 31 (2007), S. 509–533

[CPL98] Carley, K. ; Prietula, M. J. ; Lin, Z.: Design versus cognition interaction of agent cognition and organizational design on organizational performance. In: *Journal of Artificial Societies and Social Simulation* 1 (1998)

[Cre97] Crestani, F.: Application of spreading activation techniques in information retrieval. In: *Artificial Intelligence Review* 11 (1997), Nr. 6, S. 453–482

[Dam94] Damasio, A. R.: Descartes' error and the future of human life. In: *Scientific American* 271 (1994), Nr. 4

[Dam03] Damasio, A. R.: *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. Harvest Books, 2003. – ISBN 0–15–100557–5

[Dam12] Damasio, A.: In: *Psychoanalytic Review* 99 (2012), Nr. 4, S. 591–594

[Dar98] Darwin, C.: *The Expressions of the Emotions in Man and Animals*. Oxford University Press, New York, 1998

[DBD+15] Dietrich, Dietmar ; Brandster, Christian ; Doblhammer, Klaus ; Fittner, Martin ; Fodor, Georg ; Gelbard, Friedrich ; Huber, Matthias ; Matthias Jakube and, Stefan K. ; Kowarik, Daniela ; Schaat, Samer ; Wendt, Alexander ; Widholm, Roman: Natural Scientific, Psychoanalytical Model of the Psyche. Scientific Report III / TU Wien, Institute of Computer Technology. 2015. – Forschungsbericht

[Det07] Detel, W.: *Erkenntnis- und Wissenschaftstheorie*. Reclam, Stuttgart, 2007

[Deu11] Deutsch, T.: *Human Bionically Inspired Autonomous Agents. The Framework Implementation ARSi11 of the Psychoanalytical Entity Id Applied to Embodied Agents*, Vienna University of Technology, Dissertation thesis, 2011

[DFZB09] Dietrich, D. ; Fodor, G. ; Zucker, G. ; Bruckner, D.: *Simulating the Mind - A Technical Neuropsychoanalytical Approach*. Springer, Wien, 2009. – 436 S. – ISBN 9781412949644

[DG13] Drner, D ; Gss, C. D.: PSI: A Computational Architecture of Cognition, Motivation, and Emotion. In: *Review of General Psychology* 17 (2013), Nr. 3, S. 297–317

[DHK+07] Davidsson, P. ; Holmgren, J. ; Kyhlb, H. ; Mengistu, D. ; Persson, M.: Applications of agent based simulation. In: *Proceedings of the 2006 international conference on Multi-agent-based simulation VII*, 2007, S. 15–27

[DLD+01] Drner, D. ; Levi, P. ; Detje, F. ; Becht, M. ; Lippold, D.: Der agentenorientierte, sozionische Ansatz mit PSI / Universitamberg. 2001. – Forschungsbericht

[DMP07] Dessalles, J.L ; Mller, J.P. ; Phan, D.: Emergence in multi-agent systems: conceptual and methodological issues. In: Phan, D. (Hrsg.) ; Amblard, F. (Hrsg.): *Agent-based modelling and simulation in the social and human sciences*. Bardwell-

Press, Oxford, UK, 2007, S. 327–355

[DMP10] Dias, J. ; Mascarenhas, S. ; Paiva, A.: FAtiMA Modular: Towards an Agent Architecture with a Generic Appraisal Framework. (2010)

[Doe99] Doerner, D.: *Bauplan fr eine Seele*. Rowohlt, Reinbeck bei Hamburg, 1999

[DOP08] Duch, W. ; Oentaryo, R. J. ; Pasquier, M.: Cognitive Architectures: Where do we go from here? In: *AGI* 171 (2008), S. 122–136

[DP05] Dias, J. ; Paiva, A.: Feeling and Reasoning: A Computational Model for Emotional Characters. In: *Portuguese Conference on Artificial Intelligence*, 2005, S. 127–140

[DS00] Dietrich, D. ; Sauter, T.: Evolution potentials for fieldbus systems. In: *IEEE International Workshop on Factory Communication Systems WFCS 2000*, 2000

[DSB+13a] Dietrich, D. ; Schaat, S. ; Bruckner, D. ; Doblhammer, K. ; Fodor, G.: The Current State of Psychoanalytically-Inspired AI. A Holistic and Unitary Model of Human Psychic Processes. In: *Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society*, 2013

[DSB+13b] Dietrich, D. ; Schaat, S. ; Bruckner, D. ; Doblhammer, K. ; Fodor, G.: The Current State of Psychoanalytically-Inspired AIThe Current State of Psychoanalytically-Inspired AI. In: *Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society*, 2013

[DSS+16] Dietrich, D. ; Schaat, S. ; Sauter, T. ; Doblhammer, K. ; Jakubec, M. ; Widholm, R.: Funktionale Abbildung der menschlichen Psyche fr die Automatisierungs- und Regelungstechnik. In: *e&i Elektrotechnik und Informationstechnik* 133 (2016), Nr. 1, S. 48–51

[Ekm99] Ekman, P.: Basic emotions. In: *Handbook of Cognition and Emotion*. John Wiley & Sons Ltd, West Sussex, England, 1999, S. 45–60

[Eps99] Epstein, J.: Agent-Based Computational Models And Generative Social Science. In: *Complexity* 4 (1999), Nr. 5, S. 41–60

[Eps08] Epstein, J.: Why Model? In: *Journal of Artificial Societies and Social Simulation* 11 (2008), Nr. 4

[FMS07] Fum, D. ; Missier, F. D. ; Stocco, A.: The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. In: *Cognitive Systems Research* 8 (2007), Nr. 3, S. 135–142

[FR06] Franklin, S. ; Ramamurthy, U.: Motivations, Values and Emotions: 3 sides of the same coin. In: *Proceedings of the Sixth International Workshop on Epigenetic Robotics* Lund University Cognitive Studies, 2006

[Fra97] Franklin, S.: Autonomous Agents as Embodied AI. In: *Cybernetics and Systems* 28 (1997), Nr. 6

[Fri08] Frijda, N.: The Psychologists' Point of View. In: *Handbook of Emotions*. The Guilford Press, New York, 2008

[Gar86] Garcia, E. E.: Psychoanalysis: Science or fiction? In: *Jefferson Journal of Psychiatry* 4 (1986), Nr. 1, S. 7–12

[GG96] Gigerenzer, G. ; Goldstein, D.: Reasoning the fast and frugal way: Models of bounded rationality. In: *Psychological Review* 103 (1996), Nr. 4, S. 650–669

[Gig08] Gigerenzer, G.: *Gut Feeling. The Intelligence of the Unconscious*. Pinguin, New York, 2008

[Gig13] Gigerenzer, G.: Smart Heuristics. In: Brockman, John (Hrsg.): *Thinking. The Science of Decision-Making, Problem-Solving, and Predicition*. Harper, New York, 2013

[Gil06] Gilbert, N.: When Does Social Simulation Need Cognitive Models? In: Sun, R

(Hrsg.): *Cognition and Multi-Agent Interaction*. Cambridge University Press, New York, 2006

[Gil07] Gilbert, N.: Computational Social Science: Agent-based social simulation. In: *Agent-based Modelling and Simulation*. Bardwell, Oxford, 2007, S. 115 – 134.

[GM14] Gratch, J. ; Marcella, S. ; Gratch, J. (Hrsg.) ; Marcella, S. (Hrsg.): *Social Emotions in Nature and Artifact*. Oxford University Press, New York, 2014

[Gol06] Goldman, A. I.: *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press, 2006

[GT05] Gilbert, N. ; Troitzsch, K. G.: *Simulation for the social scientist*. Open University Press, Berkshire, UK, 2005

[HD09] Hagg, J. ; Drner, D.: The drunken mice. In: *Proceedings of the Ninth International Conference on Cognitive Modeling - ICCM*, 2009

[Hic09] Hickok, G.: Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans. In: *J Cogn Neurosci.* 21 (2009), Nr. 7, S. 1229–1243

[Hon95] Honderich, B.: Laws, natural or scientific. In: *Oxford Companion to Philosophy*. Oxford University Press, 1995, S. 474–476

[Hul51] Hull, C.: *Essentials of Behavior*. Yale University Press, New Haven, 1951

[IBH15] Inzlicht, M. ; Bartholow, B. D. ; Hirsh, J. B.: Emotional foundations of cognitive control. In: *Trends in Cognitive Sciences* 19 (2015), S. 32–44

[IEE12] IEEE: IEEE Standard for System and Software Verification and Validation. In: *IEEE Std 1012-2012 (Revision of IEEE Std 1012-2004)*. IEEE, 2012, S. 1–223

[Iza93] Izard, C E.: Four systems for emotion activation: Cognitive and noncognitive processes. In: *Psychological Review* 100 (1993), Nr. 1, S. 68–90

[Jam84] James, W.: What is an emotion? In: *Mind* 34 (1884), S. 188–205

[JJ03] Jager, W. ; Janssen, M.: The need for and development of behaviourally realistic agents. In: *Multi-Agent Simulations II. Lecture Notes in Computer Science Volume 2581*. Springer, Berlin Heidelberg, 2003, S. 36–49

[JJV+00] Jager, W. ; Janssen, M.A. ; Vries, H.J.M D. ; Greef, J. A. D. ; Viek, C.A.J.: Behaviour in commons dilemmas: Homo economicus and Homo psychologicus in an ecological-economic model. In: *Ecological Economics* 35 (2000), Nr. 3, S. 357–380

[Jl02] Janssen, M. A. ; lJager, W.: Stimulating diffusion of green products - Co-evolution be-tween firms and consumers. In: *Journal of Evolutionary Economics* 12 (2002), Nr. 3, S. 283–306

[Kah11] Kahneman, D.: *Thinking, fast and slow*. Penguin, London, 2011

[Kah13] Kahneman, D.: The Marvels and the Flaws of Intuitive Thinking. In: Brockman, John (Hrsg.): *Thinking. The New Science of Decision-Making, Problem-Solving, and Prediction*. Harper, New York, 2013

[Kan99] Kandel, E. R.: Biology and the future of psychoanalysis: A new intellectual framework for psychiatry revisited. In: *The American Journal of Psychiatry* (1999), S. 505–524

[Key13] Keysers, C.: *Das empathische Gehirn*. Bertelsmann, Mnchen, 2013

[KK13] Kennedy, W. G. ; Krueger, F.: Building a Cognitive Model of Social Trust Within ACT-R. In: *Proceedings of AAAI Spring Symposium: Trust and Autonomous Systems*, 2013, S. 25–27

[KL13] Kilner, J.M. ; Lemon, R.N.: What We Know Currently about Mirror Neurons. In: *Current Biology* 23 (2013), Nr. 23, S. 1057–1062

[KSSW16] Kollmann, S. ; Siafara, L. ; S.Schaat ; Wendt, A.: Towards a Cognitive Multi-Agent System for Building Control. In: *Proceedings of the Conference on Brain*

Inspired Cognitive Architectures, 2016

[Lak70] Lakatos, I.: Falsification and the methodology of scientific research programs. In: *Criticism ans the Growth of Knowledge*. Cambridge University Press, Cambridge, UK, 1970, S. 91–196

[LHU09] Leye, S. ; Himmelspach, J. ; Uhrmacher, A. M.: A discussion on experimental model validation. In: *11th International Conference on Computer Modelling and Simulation*, 2009, S. 161–167

[Lin85] Lindenberg, S.: An assessment of the new political economy: Its potential for the social sciences and for sociology in particular. In: *Sociological Theory* 3 (1985), Nr. 1, S. 99–114

[Lis09] List, E.: *Psychoanalyse*. UTB, Stuttgart, 2009

[LL03] Loewenstein, G. ; Lerner, J. S.: The role of affect in decision making. In: *Handbook of affective science*. Oxford University Press, 2003, S. 619–642

[Llo01] Lloyd, S.: Measures of complexity: a nonexhaustive list. In: *IEEE Control Systems Magazine* 21 (2001), S. 7–8

[LLR09] Langley, P. ; Laird, J. E. ; Rogers, S.: Cognitive architectures: Research issues and challenges. In: *Cognitive Systems Research* 10 (2009), S. 141–160

[Luh11] Luhmann, N.: *Einfhrung in die Systemtheorie*. Carl Auer Verlag, 2011

[Mac90] MacLean, P. D.: *The triune brain in evolution: Role in paleocerebral functions*. Springer Science & Business Media, NY, 1990

[McC02] McComas, W. F.: The Principal Elements of the Nature of Science: Dispelling the Myth. In: *The Nature of Science in Science Education*. Kluwer, NY, 2002, S. 53–72

[McI96] McIntosh, D. N.: Facial Feedback Hypotheses: Evidence, implications, and directions. In: *Motivation and Emotion* 20 (1996), Nr. 2, S. 121–147

[Mea72] Meadows, D. H.: *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*. Earth Island, London, UK, 1972

[Meg14] Megill, J.: Emotion, Cognition and Artificial Intelligence. In: *Minds & Machines* 24 (2014), S. 189–199

[Meh96] Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. In: *Current Psychology* 14 (1996), S. 261–292

[MG09] Marsella, S. C. ; Gratch, J.: EMA: A process model of appraisal dynamics. In: *Cognitive Systems Research* 10 (2009), Nr. 1, S. 70–90

[MG14] Marsella, S. ; Gratch, J.: Requirement for a Process Model of Appraisal From a Social Function Perspective. In: Gratch, J. (Hrsg.) ; Marsella, S. (Hrsg.): *Social Emotions in Nature and Artifacts*. Oxford University Press, NY, US, 2014, S. 55–69

[MGP10] Marsella, S. ; Gratch, J. ; Petta, P.: Computational models of emotion. In: Scherer, K. R. (Hrsg.): *A Blueprint for Affective Computing-A sourcebook and manual*. Oxford University Press, New York, 2010, S. 21–46.

[Mit09] Mitchell: *Complexity*. Oxford University Press, New York, 2009

[Mor03] Morrison, J. E.: A review of computer-based human behavior representations and their relation to military simulations. In: *Institute for Defense Analyses, Tech. Rep. IDA Paper P-3845*, 2003

[MR74] Mehrabian, A. ; Russell, J. A.: *An approach to environmental psychology*. MIT Press, Cambridge, 1974

[Mur02] Murphy, G. L.: *The Big Book of Concepts*. MIT Press, Cambridge, Massachusetts, 2002

[MV87] Maturana, H. R. ; Varela, F.: *The tree of knowledge: The biological roots of*

human understanding. New Science Librar, Boston, MA, 1987

[NB14] Natalini, D. ; Bravo, G.: Encouraging Sustainable Transport Choices in American Households: Results from an Empirically Grounded Agent-Based Model. In: *Sustainability* 6 (2014), Nr. 1, S. 50–69

[NBW+05] Niedenthal, Paula M. ; Barsalou, Lawrence W. ; Winkielman, Piotr ; Krauth-Gruber, Silvia ; Ric, Frans: Embodiment in Attitudes, Social Perception, and Emotion. In: *Personality and Social Psychology Review* 9 (2005), Nr. 3, S. 184–211

[New73] Newell, Alan: You can't play 20 questions with nature and win. In: *Visual Information Processing* (1973)

[New80] Newell, Allen: Physical Symbol Systems. In: *Cognitive Science* 4 (1980), S. 135–183

[NS06a] Naveh, Isaac ; Sun, Ron: A cogntively based simulation of academic science. In: *Computational and Mathematical Organzation Theory* 12 (2006), S. 313–337

[NS06b] Naveh, Isaac ; Sun, Ron: Simulating a Simple Case of Organizational Decision Making. In: Sun, Ron (Hrsg.): *Cognition and Multi-Agent Interaction*. Cambridge University Press, New York, 2006

[NSB06] Naqvi, Nasir ; Shiv, Baba ; Bechara, Antoine: The Role of Emotion in Decision Making. In: *Current Directions in Psychological Science* 15 (2006), S. 260–264

[NSS59] Newell, Allen ; Shaw, John C. ; Simon, Herbert A.: Report on a general problem-solving program. In: *FIP Congress* (1959)

[OCC90] Ortony, A. ; Clore, G. L. ; Collins, A.: *The Cognitive Structure of Emotions*. Cambridge University Press, New York, 1990

[OK07] Ouder, P. ; Kaplan, F.: What is intrinsic motivation? A typology of computational approaches. In: *Frontiers in Neurorobotics* 1 (2007), September, Nr. 6

[OSFB94] Oreskes, N. ; Shrader-Frechette, K. ; Belitz, K.: Verification, Validation, and Conformation of Numerical Models in the Earth Sciences. In: *Science* 263 (1994), S. 641–646

[Pan98] Panksepp, J.: *Affective Neuroscience, the Foundations of Human and Animal Emotions*. Oxford University Press, Inc. 198 Madison Avenue, New York, 1998. – ISBN 0195096738

[Pes08] Pessoa, L.: On the relationship between emotion and cognition. In: *Nature Reviews Neuroscience* 9 (2008), S. 148–158

[Pic97] Picard, R. W.: *Affective Computing*. MIT Press, Cambridge, 1997

[Pop59] Popper, Karl R.: *The logic of scientific discovery*. Hutchinson Education, 1959

[Pou92] Poundstone, W.: *Prisoner's dilemma*. Anchor, 1992

[PP10] Parisi, D. ; Petrosino, G.: Robots that have emotions. In: *Adaptive Behavior* 18 (2010), Nr. 6, S. 453–469

[Pyl99] Pylyshyn, Zenon: Is vision continuous with cognition? the case for cognitive impenetrability of visual perception. In: *Behavioral and Brain Sciences* 22 (1999), Nr. 3, S. 341367

[Ran09] Rank, Stefan: *Behaviour coordination for models of affective behaviour*, Technical University of Vienna, Diss., 2009

[Ras83] Rasmussen, J.: Skills, rules, knowledge; signals, signs, and symbols and other distinctions in human performance models. In: *IEEE Transactions on Systems, Man, & Cybernetics* 13 (1983), S. 257–266

[RFGF96] Rizzolatti, Giacomo ; Fadiga, Luciano ; Gallese, Vittorio ; Fogassi, Leonardo: Premotor cortex and the recognition of motor actions. In: *Cognitive Brain Research* 3 (1996), Nr. 2, S. 131–141

[RHD+13] Reisenzein, Rainer ; Hudlicka, Eva ; Dastani, Mehdi ; Gratch, Jonathan ; Hindriks, Koen ; Lorini, Emiliano ; Meyer, John-Jules: Computational Modeling of Emotion: Towards Improving the Inter- and Intradisciplinary Exchange. In: *IEEE Transactions on Affective Computing* 4 (2013), S. 246–266

[RMG86] Rumelhart, David E. ; McClelland, James L. ; Group, PDP R.: *Parallel distributed processing. Explorations in the microstructure of cognition.* MIT Press, Cambridge, MA, 1986

[RN03] Russell, S. ; Norvig, P.: *Artificial Intelligence: A Modern Approach.* Prentice Hall, New Jersey, 2003

[Rus03] Russell, James: Core affect and the psychological construction of emotion. In: *Psychological Review* 110 (2003), Nr. 1, S. 145–172

[Sal06] Salvucci, D.: Modeling driver behavior in a cognitive architecture. In: *Human Factors* 48 (2006), Nr. 2, S. 362–380

[Sar13] Sargant, R. G.: Verification and validation of simulation models. In: *Journal of Simulation* 7 (2013), S. 12–24

[SAS94] Stanley, A. ; Ashlock, D. ; Smucker, M. D.: Iterated Prisoner's Dilemma with Choice and Refusal of Partners / Iowa State University, Department of Economics. 1994. – Forschungsbericht. – 131–175 S

[SC07] Smith, E R. ; Conrey, F. R.: Agent-Based Modeling: A New Approach for Theory Building in Social Psychology. In: *Personality and Social Psychology Review* 11 (2007), S. 87–104

[Sch69] Schelling, T.: Models of Segregation. In: *The American Economic Review* 59 (1969), S. 488–493

[Sch05] Scherer, K. R.: Waht are emotions? And how can they be measured? In: *Social Science Information* 44 (2005), S. 695–729

[Sch09] Schneider, G. B.: *Wenn Agenten sich streiten. Ein Agentenmodell zur Erforschung sozialer Konflikte*, University of Kassel, Diss., 2009

[Sch12] Schaat, S.: *Integrated Drive Object Categorization in Cognitive Agents. An Activation-Based Multi-Criteria Exemplar Model*, Faculty of Electrical, TU Wien, Diplomarbeit, 2012

[Sch16a] Schaat, S.: A Case-Driven Methodology for the Interdisciplinary Development and Examination of Mental Architectures. In: *Proceedings of the Conference on Brain Inspired Cognitive Architectures*, 2016

[Sch16b] Schaat, S.: SiMA-C: A Foundational Mental Architecture. In: *Proceedings of the Conference on Brain Inspired Cognitive Architectures*, 2016

[SD14] Schaat, S. ; Dietrich, D.: Case-Driven Agent-Based Simulation for the Development and Evaluation of Cognitive Architectures. In: *Proceedings of the 26th Benelux Conference on Artificial Intelligence*, 2014, S. 73–80

[SDD14] Schaat, S. ; Dietrich, D. ; Doblhammer, K.: A Multi-level Model of Motivations and Valuations for Cognitive Agents. In: *Proceedings of the 6th International Conference on Agents and Artificial Intelligence (ICAART)*, 2014

[SDG+15] Schaat, S. ; Dickert, S. ; Geveze, E. ; Miladinovic, A. ; Wilker, S. ; Gruber, V.: Modelling Emotion and Social Norms for Consumer Simulations Exemplified in Social Media. In: *International Conference on Affective Computing and Intelligent Interaction*, 2015, S. 851–856

[SDW+13] Schaat, S. ; Doblhammer, K. ; Wendt, A. ; Gelbard, F. ; Herret, L. ; Bruckner, D.: A Psychoanalytically-Inspired Motivational and Emotional System for Autonomous Agents. In: *Proceedings of the 39th Annual Conference of the IEEE*

Industrial Electronics Society, 2013

[Sea80] Searle, J. R.: Minds, brains, and programs. In: *Behavioral and Brain Sciences* 3 (1980), S. 417– 424

[SHDD14] Schaat, S. ; Huber, M. ; Doblhammer, K. ; Dietrich, D.: An Interdisciplinary Approach for a Holistic and Embodied Emotion Model in Humanoid Agents. In: *Proceedings of the International Workshop on Artificial Intelligence and Cognition*, 2014, S. 16–26

[Sim56] Simon, H. A.: Rational Choice and the Structure of the Environment. In: *Psychological Review.* 63 (1956), Nr. 2, S. 129–138

[Sim65] Simon, H. A.: The architecture of complexity. In: *General systems* 10 (1965), S. 63–76

[Sim90] Simon, H. A.: Innvariants of human behavior. In: *Annual Review of Psychology* 41 (1990), S. 1–19

[Sim96] Simon, H.: *The Sciences of the Artificial.* MIT Press, Cambridge, 1996

[SJD16] Schaat, S. ; Jager, W. ; Dickert, S.: Psychologically Plausible Models in Agent-based Simulations of Sustainable Behavior. In: *Agent-based Models of Sustainable Behavior.* Springer Berlin Heidelberg, 2016

[SKZ+15] Schaat, S. ; Kollmann, S. ; Zhukova, O. ; Dietrich, D. ; Doblhammer, K.: Examination of Foundational AGI-Agents in Artificial-Life Simulations. In: *The 2015 Conference on Technologies and Applications of Artificial Intelligence*, 2015, S. 330–335

[SM10] Schmid, P.C. ; Mast, M.S.: Mood effects on emotion recognition. In: *Motivation and Emotion* 34 (2010), Nr. 3, S. 288–292

[SMHS09] Schuelein, J. A. ; Mikl-Horke, G. ; Simsa, R.: *Soziologie fr das Wirtschaftsstudium.* Facultas, Wien, 2009

[SMS88] Strack, R. ; Martin, L. ; Stepper, S.: Inhibiting and Facilitqating Conditions of Facial Expressions: A non-obstrusve test of the facial feedback hypotheses. In: *Journal of Personality and Social PSychology* 54 (1988), S. 768–777

[Sol96] Solms, M.: Was sind Affekte? In: *Psyche* 6 (1996), S. 487–522

[SRF00] Soucek, C. ; Russ, G. ; Fuertes, C. T.: The Smart Kitchen Project - An Application on Fieldbus Technology to Domotics. In: *Proceedings of the 2nd International Workshop on Networked Appliances (IWNA2000)*, 2000

[SS13] Senge, K. ; Schuetzeichel, R. ; Senge, K. (Hrsg.) ; Schuetzeichel, R. (Hrsg.): *Hauptwerke der Emotionsoziologie.* Springer VS, Wiesbaden, 2013

[SS14] Smart, P. R. ; Sycara, K.: Cognitive Social Simulation and Collective Sensemaking: An Approach Using the ACT-R Cognitive Architecture. In: *The Sixth International Conference on Advanced Cognitive Technologies and Applications*, 2014, S. 195–204

[Str03] Strongman, K.T.: *The Psychology of Emotion. From Everyday Life to Theory.* John Wiley & Sons Ltd, West Sussex, England, 2003

[Sun06a] Sun, R.: The CLARION Cognitive Architecture: Extending Cognitive Modeling to Social Simulation. In: *Cognition and Multi-Agent Interaction.* Cambridge University Press, New York, 2006

[Sun06b] Sun, R. ; Sun, R. (Hrsg.): *Cognition and Multi-Agent Interaction.* Cambridge University Press, New York, 2006

[Sun09] Sun, R.: Cognitive architectures and multi-agent social simulation. In: *Multi-Agent Systems for Society.* Springer Berlin Heidelberg, 2009, S. 7–21

[Sun12] Sun, R.: *Grounding social sciences in cognitive sciences.* MIT Press, MA, US, 2012

[SWB13] Schaat, S. ; Wendt, A. ; Bruckner, D.: A Multi-Criteria Exemplar Model for

Holistic Categorization in Autonomous Agents. In: *Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society*, 2013

[SWJ+14] Schaat, S. ; Wendt, A. ; Jakubec, M. ; Gelbard, F. ; Herret, L. ; Dietrich, D.: ARS: An AGI Architecture. In: *Proceedings of the 7th conference on Artificial General Intelligence, LNAIC*, Springer Switzerland, 2014, S. 155–164

[SWK+15] Schaat, S. ; Wendt, A. ; Kollmann, S. ; Gelbard, F. ; Jakubec, M.: Interdisciplinary Development and Evaluation of Cognitive Architectures Exemplified with the SiMA Approach. In: *EuroAsianPacific Joint Conference on Cognitive Science*, 2015, S. 515–520

[TK74] Tversky, A. ; Kahneman, D.: Judgement Under Uncertainty: Heuristics and Biases. In: *Science* 185 (1974), S. 1124–1131

[Tom99] Tomasello, M.: *The Cultural Origins of Human Cognition.* Harvard University Press, 1999

[Tom08] Tomasello, M.: *Why we cooperate.* The MIT Press, Cambridge, Mass., 2008

[TS09] Thaler, R. ; Sunstein, C.: *Nudge.* Pinguin Books, London, 2009

[Vel08] Velik, R.: *A Bionic Model for Human-like Machine Perception*, Vienna University of Technology, Institute of Computer Technology, Diss., 2008

[Via00] Viale, R.: The Mind-Society Problem. In: *Mind & Society* 1 (2000), S. 3–24

[Via12] Viale, R.: *Methodological Cognitivism: Vol. 1: Mind, Rationality, and Society.* Springer, Berlin., 2012

[VS06] de Vignemont, F ; Singer, T.: The empathic brain: How, when, and why? In: *Trends in Cognitive Science* 10 (2006), S. 435–441

[Waa12] Kap. Empathy in Primates and Other Mammals In: de Waal, F.: *Empathy. From Bench to Bedside.* MIT Press, Cambridge, MA, 2012, S. 87–102

[War97] Warsitz, P.: Die widerstige Erfahrung der Psychoanalyse zwischen den Methodologien der Wissenschaften. In: *Psyche* 51 (1997), S. 101–142

[Wen05] Wenstop, F.: Mindset, Rationality and Emotion in Multi-criteria Decision Analysis. In: *Journals of Multi-Criteria Decision Analysis* 13 (2005), S. 161–172

[WGF+15] Wendt, A. ; Gelbard, F. ; Fittner, M. ; Schaat, S. ; Jakubec, M.: Decision-Making in the Cognitive Architecture SiMA. In: *The 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI 2015)*, 2015

[Wie10] Wiest, G.: *Hierarchien in Gehirn, Geist und Verhalten: ein Prinzip neuraler und mentaler Funktion.* Springer-Verlag, Wien, 2010

[WSG+13] Wendt, A. ; Schaat, S. ; Gelbard, F. ; Muchitsch, C. ; Bruckner, D.: Usage of spreading activation for content retrieval in an autonomous agent. In: *Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Societ*, 2013

[Wun22] Wundt, W.: *Vorlesung ber die Menschen- und Tierseele.* Voss Verlag, Leipzig, 1922

[Zei10] Zeilinger, H.: *Bionically Inspired Information Representation for Embodied Software Agents*, Institute of Computer Technnology, TU Wien, Diss., 2010

Curriculum Vitae

I started my exploration of the underlying mechanisms of human behavior by studying Medicine. After getting an overview of basic human anatomy and physiology, I switched my studies to Medical Informatics for a systematic and systemic methodology - conducted at a higher level of representation (i.e., at the mental level) using a top-down approach. Following an interdisciplinary approach, I focused on Cognitive Science and Computer Simulation. Additional to developing methodologies to bridge scientific and technical disciplines, I aim to support bridging different levels of explanation. I think different aspects of a system should be addressed and integrated with different criteria. In particular, I am working on dualistic models of decision making that ground reasoned decisions on desires and emotion, going beyond the limits of the classical rational choice approach. Therefore, I am interested in developing computational models of adaptive decision making that operationalize the underlying psychological mechanisms of heuristics. Furthermore, I am interested in using computational models to base social behavior on the individual level, with consideration of the impact of social context on decision making. Here, I am especially interested in balancing reduction against specification of decision models in concrete socially relevant applications. This includes any application, where simulations of human behavior have to be considered. Examples are: the development and evaluation of policies that aim to enforce specific behavior, the development of technical artifacts to apply these policies, and user simulations for enhancing socio-technical integration. Another line of research interest considers the translation of insights from research of the human control system into artificial control systems.

Personal Information

Birthday: 19.10.1982
Place of birth: Vienna, Austria
Email address: s.schaat@gmail.com

Research Experience

- Project Leader and Researcher, Institute of Computer Technology, TU Wien, Austria, since 2011
- Cognitive building automation

  - Co-PI (principal investigator) of project KORE, since 10/2015
  - Co-PI of project ECABA, 09/2014 – 09/2015

- Cognitive simulation of environmental-friendly decisions,
  PI and consortial leader of project CogMAS, 10/2014 – 10/2015

- Cognitive architectures

  - PI of project SiMA, 05/2013-12/2015
  - Project assistant, project ARS, 03/2012 – 04/2013

- Adaptive street lighting, Project assistant, project SIRIUS, 08/2012-04/2014

- Semantic web technologies, Project assistant, project VKT GÖPL, 07/2011-02/2012

- Interoperability and data integration, Research Assistant,
  Institute for medical information systems, Medical University Vienna 04/2010-06/2011

University Education

- Ph.D, Simulation of Foundational Human Information-Processing in Social Context, 11/2012 - 08/2016

  - Institute of Computer Technology, Vienna University of Technology (TU Wien)
  - Supervisor: o.Univ.Prof. Dipl.-Ing. Dr. Dietmar Dietrich
  - Co-Supervisor: Prof. Dr. Cristiano Castelfranchi

- B.Sc and M.Sc., Medical Informatics, 2010-2012

  - Vienna University of Technology (TU Wien) and University of Vienna, Austria
  - Thesis: "Integrated Drive Object Categorization in Cognitive Agents" at Institute of Computer Technology (Prof. Dietmar Dietrich)
  - Course abroad: EU-Athens program, course „Collective Intelligence" at Telecom Paristech

- Medicine, 2000-2004

  - University of Vienna, Austria
  - Not completed (only non-clinical courses completed)

Peer Reviewed Publications

1. S. Schaat, "SiMA-C: A Foundational Mental Architecture", Proceedings of the Conference on Brain Inspired Cognitive Architectures, July 2016, in press.

2. S. Schaat, "A Case-Driven Methodology for the Interdisciplinary Development and Examination of Mental Architectures", Proceedings of the Conference on Brain Inspired Cognitive Architectures, July 2016, in press.

3. S. Kollmann, L. Siafara, S.Schaat, A. Wendt, „Towards a Cognitive Multi-Agent System for Building Control“, Proceedings of the Conference on Brain Inspired Cognitive Architectures, July 2016, in press.

4. S. Schaat, W. Jager, S. Dickert, A. Miladinovic, “Psychologically Plausible Models in Agent-based Simulations of Sustainable Behavior“, in Agent-Based Modeling of Sustainable Behaviors, (eds.) A. Alonso-Betanzos, N. Sánchez-Maroño, O. Fontenla-Romero, J. Bajo, J.A. Corchado, in press.

5. D. Dietrich, M. Jakubec, S. Schaat, K. Doblhammer, G. Fodor, C. Brandstaetter, “The Fourth Outrage of Man - Is the Turing-test Still Up to Date?”, Journal of Computers 12(2), pp. 116–126, in press.

6. C. Brandstaetter, S. Schaat, A. Wendt, M. Fittner, “How Agents use Breadcrumbs to Find their Way”, Journal of Computers 12(1), pp. 89–96, in press.

7. D. Dietrich, S. Schaat, T. Sauter, K. Doblhammer, M. Jakubec, R. Widholm. „Funktionale Abbildung der menschlichen Psyche für die Automatisierungs- und Regelungstechnik“. ei Elektrotechnik und Informationstechnik, 133(1), 48-51, 2016.

8. D. Dietrich. D. Bruckner, K. Doblhammer, S. Schaat, „Psychoanalyse und Computertechnik, Psychoanalytiker und Ingenieure“, Prostir M GmbH, Ukraine, Lviv, Jahrgang – 2015, April, Band N9; p. 97-109 (in ukrainisch) [’ . .]

9. S. Schaat, S. Kollmann, O. Zhukova, D. Dietrich, K. Doblhammer. ”Examination of Foundational AGI-Agents in Artificial-Life Simulations”. The 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI 2015), pp. 330-335, Tainan, Nov. 2015.

10. A. Wendt, F. Gelbard, M. Fittner, S. Schaat, M. Jakubec. ”Decision-Making in the Cognitive Architecture SiMA”. The 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI 2015), Tainan, 2015.

11. S. Schaat, A. Wendt, S. Kollmann, F. Gelbard, M. Jakubec. “Interdisciplinary Development and Evaluation of Cognitive Architectures Exemplified with the SiMA Approach”. EuroAsianPacific Joint Conference on Cognitive Science, pp. 515-520, Turin, Sept. 2015.

12. S. Schaat, A. Miladinović, S. Wilker, S. Kollmann, S. Dickert, E. Geveze, V. Gruber. „Emotion in Consumer Simulations for the Development and Testing of Recommendations for Marketing Strategies”. Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems, pp. 25-32, Vienna, Sept. 2015.

13. G. Zucker, U. Habib, M. Blöchle, A. Wendt, S. Schaat, L. C. Siafara, “Building Energy Management and Data Analytics”. 5th Symposium on Communications for Energy systems, Vienna, Aug.2015.

14. S. Schaat, S. Dickert, E. Geveze, A. Miladinovic, S. Wilker, V. Gruber, “Modelling Emotion and Social Norms for Consumer Simulations Exemplified in Social Media”, International Conference on Affective Computing and Intelligent Interaction, pp. 851-856, Xian, 2015.

15. D. Dietrich, S. Schaat, M. Jakubec, A. Malakova, “Applying the Methods of Data Mining and Knowledge Engineering to the SiMA Project”, ITIDS ’2015 Information Technologies for Intelligent Decision Making Support, pp. 13-16, 2015.

16. S. Schaat, M. Huber, K. Doblhammer, and D. Dietrich, "An Interdisciplinary Approach for a Holistic and Embodied Emotion Model in Humanoid Agents", International Workshop on Artificial Intelligence and Cognition, pp. 16-26, 2014/.

17. S. Schaat and D. Dietrich, "Case-Driven Agent-Based Simulation for the Development and Evaluation of Cognitive Architectures", 26th Benelux Conference on Artificial Intelligence, pp. 73-80, 2014.

18. S. Schaat, A. Wendt, M. Jakubec, F. Gelbard, L. Herret, and D. Dietrich, "ARS: An AGI Architecture", Proceedings of the 7th conference on Artificial General Intelligence, pp. 155-164, LNAIC, Springer Switzerland, 2014.

19. T. Novak, K. Pollhammer, H. Zeilinger, and S. Schaat, "Intelligent Streetlight Management in a Smart City", Proceedings of the 19th IEEE International Conference on Emerging Technologies and Factory Automation, Barcelona, 2014.

20. S. Schaat, D. Dietrich, and K. Doblhammer, "A Multi-level Model of Motivations and Valuations for Cognitive Agents", in Proceedings of the 6th International Conference on Agents and Artificial Intelligence (ICAART), pp. 255-261, Angers, 2014.

21. D. Dietrich, S. Schaat, D. Bruckner, K. Doblhammer and G. Fodor, "The Current State of Psychoanalytically-Inspired AI", in Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, 2013.

22. S. Schaat, A. Wendt, and D. Bruckner, "A Multi-Criteria Exemplar Model for Holistic Categorization in Autonomous Agents", in Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, 2013.

23. S. Schaat, K. Doblhammer, A. Wendt, F. Gelbard, L. Herret, and D. Bruckner, "A Psychoanalytically-Inspired Motivational and Emotional System for Autonomous Agents, in Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, 2013.

24. D. Bruckner, F. Gelbard, S. Schaat, and A. Wendt, „Validation of cognitive architectures by use cases," in Proceedings of the International Symposium on Industrial Electronics, Montreal, 2013.

25. S. Schaat, A. Wendt and D. Bruckner, „Integrating top-down and bottom-up perception in a cognitive architecture", abstract, Proceedings of the annual meeting of the Cognitive Science society, Berlin, 2013.

26. A. Wendt, S. Schaat, F. Gelbard, C. Muchitsch, and D. Bruckner, „Usage of spreading activation for content retrieval in an autonomous agent," in Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, 2013.

27. T. Novak, H. Zeilinger, and S. Schaat, „Increasing Energy Efficiency with Traffic Adapted Intelligent Streetlight Management", in Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, 2013.

Conference Organization and Review Service

- Organization of the two-day workshop "Neuro-psychoanalysis and computer technology" with invited guest Prof. Mark Solms, 2015

- Programm comitee Artificial General Intelligence conference, 2015, 2016
- Special session chair IECON 2013 "Cognitive Architectures and Multi-Agent Systems"
- Reviews for Books (Agent-base Models for Sustainable Behavior, 2016), AGI (2015-2016), Frontiers in Psychology (2015), IECON (2014-2016)
- Presentation of published papers (see first-author papers above) at conferences

Teaching Experience

- Supervision of seminar (6), bachelor (4), and master theses (2)
- Lectures about AI, Cognitive Science, and the SiMA Project at courses thought at the Institute of Computer Technology, TU Wien:
  - Seminar Distributed Systems, 2013-2014
  - Networked Systems, 2013-2014
  - Computer Technology Specialization, 2013-2015
  - Ausgerechnet Elektrotechnik! 2015

Press Coverage(in German)

- Die Presse: „Wie bringt man Maschinen Emotion und Intuition bei?" [How to bring emotion and intuition into machines?], Jan. 2015, Website
- Wiener Zeitung: „Warum Roboter keinen Schmäh' haben" [Why robots have no humor], Jan. 2015, Website
- Kurier: „Psychologie trifft Computertechnik" [Psychology meets computer technology], March 2015, only print.
- Interview for Ö1 (Austrian radio station) reused for SWR2 (German radio station): „Norbert Wiener – Begründer der Kybernetik als Universalwissenschaft für Maschine und Mensch" [Norbert Wiener – founder of cybernetics as a universal science for machines and humans], March 2014

Non-academic Work Experience

- 02/2009-11/2009: Tangram Multikulturelles Netzwerk, Software Engineering: Planning and development of a new administration software for a social organization (funded by the City of Vienna).
- 12/2007-05/2008: Webdirect, Webprogramming and Webdesign.
- Different jobs in school time/early student life (gastronomy, warehouse, post office, sales, customer care, technical support)

Language Skills

- German (native language)
- English (fluent in speaking and writing)
- Arabic (mediate level in speaking)
- Italian (basic knowledge in speaking and writing)

School Education

- Highschool (Naturwissenschaftliches Realgymnasium), Laaer-Berg-Strasse, 1100 Vienna, 1992-2000
- Elementary school, Schrankenberggasse , 1100 Vienna, 1988-1992