Benjamin Strenge

Computational Analysis of Task-Related Mental Representation Structures for User-Adaptive Cognitive Assistance Systems

Bielefeld University Library – PUB Theses

Strenge, Benjamin, 2020
Computational analysis of task-related mental representation structures for user-adaptive cognitive assistance systems
Bielefeld University
Germany

COMPUTATIONAL ANALYSIS OF TASK-RELATED MENTAL REPRESENTATION STRUCTURES FOR USER-ADAPTIVE COGNITIVE ASSISTANCE SYSTEMS

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)

an der Fakultät für Psychologie und Sportwissenschaft
basierend auf Forschung am
Center of Excellence in Cognitive Interaction Technology
(CITEC)
der
Universität Bielefeld

von
BENJAMIN STRENGE

Gutachter:

Prof. Dr. Thomas Schack
Jun.-Prof. Dr. Christoph Schütz

Tag der Disputation: 15. Dez. 2020

Zusammenfassung

Die Vorhersage menschlicher Fehler in Handlungssequenzen ist für zahlreiche private und berufliche Aktivitäten von immensem Wert. Bei gefährlichen oder anderweitig kritischen Aktivitäten, beispielsweise wenn falsche Aktionen irreversibel wären, ist es von entscheidender Bedeutung, Handlungsfehler zu antizipieren und zu verhindern. Doch auch bei weniger kritischen Aktivitäten können entsprechende Vorhersagen dazu dienen, Fehler zu verhindern und Handlungen somit reibungsloser und effizienter durchzuführen, als wenn irrtümlich ausgeführte Aktionen nachträglich korrigiert werden müssten. Wesentliche Bedeutung kommt der Einschätzung individueller aufgabenbezogener Vorkenntnisse auch im Bereich technischer Assistenzsysteme zu. So ist insbesondere bei der Verwendung von Datenbrillen, die durch „Erweiterte Realität“ (Augmented Reality) virtuelle Elemente direkt in das Sichtfeld der Nutzer einblenden und die Wahrnehmung der natürlichen Umgebung überlagern, eine weitreichende Schonung der Aufmerksamkeit und anderer begrenzter kognitiver Ressourcen des Nutzers unabdingbar. Ein solches System sollte also hinreichende Kenntnis des tatsächlichen Assistenzbedarfs haben und durch entsprechend gezielte Unterstützung sinnvolle kognitive Assistenz anbieten, statt Nutzer mit unnötigen Einblendungen abzulenken. Andernfalls ist mit Überforderung, verringerter Gebrauchstauglichkeit und unzureichender Nutzerakzeptanz zu rechnen.

Die strukturdimensionale Analyse mentaler Repräsentationen (SDA-M) ist ein ursprünglich aus der Kognitionspsychologie stammendes Verfahren, das sich inzwischen auch in der Bewegungs- und Sportwissenschaft und der kognitiven Robotik etabliert hat. Anhand eines speziellen, teilautomatischen Befragungsverfahrens, der sogenannten „Splitprozedur“, ermittelt SDA-M Daten über die individuellen aufgabenbezogenen Gedächtnisstrukturen einer bestimmten Person. Diese Daten wurden bisher zumeist mittels eines hierarchischen Clusteringverfahrens analysiert und in Form von Dendrogrammen visualisiert. Diese Dendrogramme und dazugehörige statistische Größen können von entsprechend geschulten Experten ausgewertet werden, um mögliche Probleme bzgl. der Handlungsausführung

zu identifizieren und Verbesserungsvorschläge zu entwickeln, bspw. im Bereich des Trainings, manueller Handlungen und der klinischen Rehabilitation. Dieses Verfahren erfordert jedoch spezielle Expertise und kostet Zeit.

In der vorliegenden Arbeit wurde daher untersucht, wie die Analyse aufgabenbezogener mentaler Repräsentationsstrukturen weiter automatisiert werden kann. Hierzu wurden verschiedene algorithmische Ansätze entwickelt, die auf unterschiedlichen kognitiven Architekturmodellen basieren. Insgesamt vier empirische Studien haben diese algorithmischen Ansätze zur Vorhersage der aufgabenbezogenen Gesamtkompetenz und der individuellen Wahrscheinlichkeiten von Fehlern bei einzelnen Aktionen in Handlungssequenzen sowohl mit dem traditionellen Experten-basierten Ansatz verglichen, als auch in unterschiedlichen praktischen Anwendungsbereichen evaluiert. Hierzu gehören eine kontrollierte Laborstudie mit einer Standardaufgabe für Zusammenbauprozesse, eine Bewegungssequenz aus dem traditionellen Kampfkunsttraining, sowie eine manuelle Montageaufgabe im angewandten industriellen Kontext.

Die empirische Evidenz belegt, dass die neuen computergestützten Analyseverfahren dem bisherigen Experten-basierten Ansatz mindestens ebenbürtig sind und die Mehrzahl der tatsächlich aufgetretenen menschlichen Fehler in den verschiedenen Anwendungsbereichen korrekt vorhersagen konnten. Die Genauigkeit der algorithmischen Vorhersagen war zudem in sämtlichen Untersuchungen signifikant über dem Zufallsniveau.

Ergänzend zu diesen theoretischen und empirischen Arbeiten im Bereich der Kognitionswissenschaft beschreibt die vorliegende Arbeit eine neue Methodik und ein entsprechendes iteratives Prozessmodell, um bei der agilen Entwicklung nutzeradaptiver Assistenzsysteme sowohl den übergeordneten, intendierten Mehrwert des Systems, als auch ethische Aspekte und relevante Eigenschaften der Stakeholder systematisch und explizit in die Systemgestaltung einzubeziehen. Diese Methodik wurde speziell entwickelt, um die Erfolgsaussichten bei der Gestaltung und Einführung zukünftiger nutzeradaptiver kognitiver Assistenzsysteme zu optimieren.

Summary

Predicting human error in action sequences is of immense value in numerous private and professional activities. In the case of dangerous or otherwise critical activities, for example if wrong actions would be irreversible, it is of crucial importance to anticipate and prevent mistakes. However, even in the case of less critical activities, corresponding predictions can serve to prevent errors and thus to carry out actions more smoothly and efficiently than if erroneously executed actions had to be corrected afterwards. The assessment of individual, task-related prior knowledge is also highly important in the area of technical assistance systems. Especially when using augmented reality smart glasses, which enrich the natural perception of environments by projecting virtual elements directly into the user's field of view, it is essential to spare users' attentional and other limited cognitive resources. Such a system should therefore have sufficient knowledge of the actual need for assistance and offer meaningful cognitive assistance through appropriately targeted support, instead of distracting users with unnecessary overlays. Otherwise, cognitive overload would reduce the system's usability and lead to a lack of user acceptance.

The structural-dimensional analysis of mental representations (SDA-M) is a method originating from cognitive psychology, which has been established in movement and sports science, as well as in cognitive robotics. Using a special, semi-automatic survey, the so-called "split procedure", SDA-M retrieves data about the individual, task-related memory structures of a particular person. So far, this data has commonly been analyzed using hierarchical clustering and visualized in the form of dendrograms. These dendrograms and the associated statistical quantities can be evaluated by appropriately trained experts in order to assess problems related to the execution of actions and to develop suggestions for improvement, for example in the area of training, manual actions, and clinical rehabilitation. However, this procedure requires special expertise and time.

The present work therefore examined how the analysis of task-related mental representation structures can be further automated. Various algorithmic approaches based on different cognitive architecture models were de-

veloped for this purpose. A total of four empirical studies have evaluated these algorithmic approaches for predicting the overall task-related competence and the individual probabilities of errors for each action in different action sequences in comparison with the traditional expert-based approach, as well as in different practical areas of application. This includes a controlled laboratory study with a standard task for assembly processes, a movement sequence from traditional martial arts training, as well as a manual assembly task in an applied industrial context.

The empirical evidence shows that the new computer-aided analysis methods are at least equal to the previous expert-based approach and that they were able to correctly predict the majority of the actual human errors in the various areas of application. The accuracy of the algorithmic predictions was also significantly above chance level in all studies.

Complementary to this theoretical and empirical work in the field of cognitive science, the present work describes a new agile methodology and a corresponding iterative process model, which systematically analyzes and considers the intended worth and outcomes of system usage, ethical issues, and relevant stakeholder characteristics during the design and development of technical systems. This methodology was specially developed to optimize the chances that advanced user-adaptive cognitive assistance systems turn out valuable and successful.

Acknowledgments

I would like to thank *all my relations* but especially...

... **Thomas Schack** for providing me with ample resources and liberties for this research and introducing me to eclectic perspectives on nature.

... **Karla** and **Manfred Strenge** for everything, but especially for believing in me and my capabilities at all times.

... **Martina Brunsmann** for being an amazing person and taking great care of our daughter Nora and other common issues when I was writing this.

... **Nora** for making me incredibly proud of her every day.

... **Kai Essig** for managing Project ADAMAAS and providing a role model for how to stay calm and serene in face of demanding situations.

... **Karsten Nebe**, **Holger Fischer** and **Gerd Szwillus** for inspiring me to join the exciting field of human-machine interaction research.

... **Ludwig Vogel** for expediting our study of scientific confrontation between human experts and algorithms.

... **Alexander Neumann** for being the most competent and humorous colleague one could ask for.

... **Christoph Schütz** for reviewing this thesis and, like **Dirk Koester** and many other smart people at CITEC, for sharing their knowledge and experience.

... **Sergej, Nick, Sebastian, Rachel, Jens** and **Simon** for enriching my existence in various precious ways.

Contents

The following chapters of this thesis are based on manuscripts that have been published or submitted for publication:

Chapter 3, 4 and **6** are based on parts of

Strenge, B., Vogel, L., & Schack, T. (2019).

Computational assessment of long-term memory structures from SDA-M related to action sequences. *PLOS ONE, 14*(2), e0212414. Public Library of Science. doi:10.1371/journal.pone.02124140

Author contributions: Conceptualization: BS, TS. Formal analysis: BS. Funding acquisition: TS. Investigation: BS, LV. Methodology: BS, LV, TS. Project administration: TS. Resources: LV. Software: BS. Supervision: TS. Validation: BS, TS. Visualization: BS. Writing – original draft: BS, LV, TS. Writing – review & editing: BS, TS.

Chapter 7 is based on

Strenge, B., & Schack, T. (submitted to *Scientific Reports*).

Empirical relationships between algorithmic SDA-M-based memory assessments and human errors in manual assembly tasks.

Author contributions: BS, TS conceived the studies. BS conducted the studies and analysed the results. BS, TS reviewed the manuscript.

Chapter 8 is based on

Strenge, B., Koester, D., & Schack, T. (2020).

Cognitive Interaction Technology in Sport – Improving Performance by Individualized Diagnostics and Error Prediction. *Frontiers in Psychology, 11*, 3641. doi:10.3389/fpsyg.2020.597913

Author contributions: BS was either solely accountable for, or involved in, all aspects of this work. DK contributed to the experimental design and execution, data preparation, and writing. TS provided the theoretical framework for SDA-M, supervised the research, and contributed to the writing.

Chapter 9 is based on

Strenge, B., & Schack, T. (2019).

AWOSE - A Process Model for Incorporating Ethical Analyses in Agile Systems Engineering. *Science and Engineering Ethics, 26*(2), 851-870. Springer Nature. doi:10.1007/s11948-019-00133-z

Author contributions: Writing – original draft: BS. Writing – review & editing: BS, TS.

Section 1.2 explains the incorporation of manuscripts in more detail.

Chapter 1.

Introduction

This thesis reflects the scientific culmination of my increasingly interdisciplinary personal journey: After experiencing the adventures of a wonderful childhood, surviving numerous impending "ends of the world" (including the Y2K problem at the turn of the millennium), attaining a brown belt in Shaolin Karate, conquering the World of Warcraft, and gaining the German general qualification for university entrance ("Abitur"), I enrolled into the computer science program at Paderborn University and chose psychology as a minor subject. This made me aware of some interesting analogies: While computer science unsurprisingly deals with information processing by computers, (cognitive) psychology investigates and models information processing by humans. The fundamental commonalities and differences between these two types of information processors have fascinated me ever since. A bit later I was thrilled to learn about the field of cognitive science, how it abstracts to some degree from the type of processor (i.e. human, machine, or other) and considers information processing by the more abstract concept of "intelligent systems". In parallel, I delved into the field of human-machine interaction research where the interests of computer science and psychology largely overlap, e.g. regarding the conception of suitable interfaces between humans and technical systems. This thesis deals with both of these areas of research. In a nutshell, it is about predictive algorithms using cognitive models for simulating some aspects of human cognition related to specific tasks based on automatized survey procedures, evaluating them in diverse fields such as industrial and sports applications, and integrating them into a larger technical system with ethical issues in mind.
However, first things first...

1.1 Basic terms

In order to do justice to an interdisciplinary readership, it seems appropriate to explain some basic terms from the relevant subject areas. This includes concepts from psychology, computer science, and neurobiology. In this section, as well as in the following chapter, notions that bear significant relevance later in this work are highlighted ***italicized and bold***, whereas less relevant terms are merely *italicized*. Unless otherwise stated by explicit references, the definitions and explanations in the following paragraph on cognitive science are derived and condensed from Anderson (2020) and serve as a courtesy toward readers outside this field.

The primary scientific foundations of this thesis could be considered as belonging mainly to the field of ***cognitive science***, which aims to integrate research from psychology, philosophy, linguistics, neuroscience, and computer science. According to the current scientific understanding, ***neurons*** are the basic units of human information processing. Neurons are cells that communicate via so-called *synapses* by releasing specific chemicals (*neurotransmitters*). Synapses are near contacts between the "outlets" of one neuron (*axon terminal boutons*) and the "inlets" of another neuron (*dendrites*). A human brain contains approximately 10^{11} neurons, many of which are active in parallel to generate the massive processing power needed to drive human behavior. Each neuron receives input signals from (on average) about 1,000 other neurons and transmits output signals to another 1,000 neurons. These input signals are accumulated on the neurons cell body (*soma*). If the accumulated input exceeds a specific threshold of a given neuron, then this neuron "fires", i.e. it transmits information by sending a nerve impulse (a so-called *action potential* or *spike*) from the soma through a long tube called *axon* towards its terminal boutons and then via synapses to other neurons. Synaptic signals can have *excitatory* or *inhibitory* effect, depending on whether their associated electrochemical activity increases or decreases the chance that the receiving neuron fires. The number of action potentials that a neuron transmits per second is called its *rate of firing*, which can be understood as its ***activation*** *level*. It is assumed that neurons represent information by responding to specific features of a stimulus. While adult humans are not presumed to

grow a substantial number of new neurons, the properties of synaptic connections are presumed to change through learning, which enables the brain to store knowledge and reproduce the associated patterns of neural activity. Cognitive science commonly abstracts the representation of knowledge by clusters of associated neurons to higher-level structures called ***chunks***, which encode facts or other concepts by connecting their associated elements (see Anderson, 2009). Analogously to information transmission among associated neurons, currently attended items (chunks) make associated memories (chunks) more available through ***spreading activation***. Factual and event-related explicit knowledge, which is usually available for conscious verbalization, is sometimes denoted as *declarative knowledge*, whereas implicit memories and skills may be attributed to *procedural knowledge*; however, this is only one of many possible distinctions that have been proposed to investigate and describe human memory systems.

The interdisciplinary field of human-machine interaction strives to apply and extend the knowledge about the human mind and behavior in the context of interaction with technical systems, e.g. in order to improve usability-related aspects. ***Usability*** is a property of technical systems, software, and other products. The international multi-part standard ISO 9241 defines usability as the *"extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* (ISO 9241-11:2018). Analogous to software engineering, ***usability engineering*** (UE) can be defined as a systematic engineering procedure for establishing this property. This term therefore primarily underlines the requirement for methodically systematic structuring of corresponding processes. The term ***user-centred design*** (UCD), which is also frequently used, primarily emphasizes the active involvement of real users in the development process (see e.g. Mao, Vredenburg, Smith, & Carey, 2001). In addition, there are various other terms such as *user-oriented design*, ***human-centered design***, ***user experience*** (UX)[1], *software ergonomics* or *human factors integration*. Each of these terms reflects different perspectives or

[1]UX is sometimes erroneously understood as a broader or more advanced concept than usability. However, according to the current ISO 9241-11 standard, the satisfaction component of usability *"includes the extent to which the user experience that results from actual use meets the user's needs and expectations"*. Therefore, the concept of usability comprises all aspects of UX that can be influenced by system designers.

focal points, but the main goal is always to achieve good usability of the system by incorporating the requirements and needs of prospective users during its development. In the context of this work, these terms can be considered as mostly synonymous.

A characteristic property of all user-centred design processes, as defined by ISO 9241-210, is their iterative organization. This arranges for *formative evaluation* activities at the end of each cycle in order to assess the currently achieved level of usability-related qualities and, if necessary, incrementally improve it. Formative evaluations can be carried out using different document- and expert-based methods, e.g. *heuristic evaluation* (Nielsen, 1993) and *cognitive walkthrough* (Polson, Lewis, Rieman, & Wharton, 1992; Wharton, Rieman, Lewis, & Polson, 1994), or ***usability tests*** (also known as *user tests*) involving observation of prospective users who try to use the system to perform a given set of tasks while an experimenter measures metrics like number of errors or task completion time.

Most of the cognitive science or usability-related methods that were developed or applied within the scope of this thesis were aimed, alongside various other goals, at improving certain aspects of ***augmented reality*** (AR) systems. In contrast to *virtual reality* (VR) systems, AR does not fully immerse users in a purely artificial environment, but it "adds" virtual elements to the normal sensory perception of natural environments. An example for AR devices are so-called ***smart glasses***, i.e. special head-mounted displays, such as the *Microsoft HoloLens*, which are worn similar to regular corrective glasses and enrich (or "augment") users' vision with two- or three-dimensional virtual elements. This technology enables displaying context-related information, e.g. as a type of assistance or instruction, directly at relevant places within users' field of view.

1.2 Sources and structure

The following doctoral thesis is divided into two parts: ***Part I*** focuses more on theoretical and empirical work in the field of cognitive science and associated areas. The research efforts reported in this part were mainly concerned with the motivation, theoretic derivation, and empiric evaluation of

Part	Chapter	Primary source
-	1	Original composition
-	2	Original composition
I	3	Strenge, B., Vogel, L., & Schack, T. (2019) in *PLOS ONE*
I	4	Strenge, B., Vogel, L., & Schack, T. (2019) in *PLOS ONE*
I	5	Original composition
I	6	Strenge, B., Vogel, L., & Schack, T. (2019) in *PLOS ONE*
I	7	Strenge, B., & Schack, T. (under review, *Scientific Reports*)
I	8	Strenge, B., Koester, D., & Schack, T. (2020) in *Frontiers in Psychology*
II	9	Strenge, B., & Schack, T. (2019) in *Science and Engineering Ethics*
II	10	Original composition
-	11	Original composition

Table 1.1: **Primary origins of chapters' contents.**

computational approaches that automate the analysis of individual mental representation structures related to action sequences. ***Part II*** focuses more on applied scientific contributions in the area of human-machine interaction. Specifically, it reports the methodical procedures by which the cognitive assessment methods from Part I could be integrated into a wearable AR-based cognitive assistance system. For this purpose, the second part introduces a new agile system development methodology that systematically takes stakeholder characteristics, intended worth and usage outcomes, usability aspects, and ethical issues into consideration. A subsequent common conclusion and outlook section reflects on both of these parts.

The thesis is mainly composed of content from four article manuscripts. Two of these have been successfully peer-reviewed and published in 2019 by *PLOS ONE* and *Science and Engineering Ethics*, another one in 2020 by *Frontiers in Psychology*. The remaining one has been under review at *Scientific Reports* as of December 2020. The first page of each chapter in the main body of the thesis contains a box that denotes the primary origin of the chapter's content. As shown in Table 1.1, most of the content of the article published in 2019 by *PLOS ONE* has been divided into three thesis chapters. All manuscripts have been revised, adjusted, and ordered to form a coherent

composition. Most evidently, the formatting of all content has been unified. This includes a common list of references at the very end of the thesis with a standardized bibliography style. All cross references between sections, paragraphs, etc. have been adjusted, and in some cases added, to match the new numbering system of the thesis. The original article manuscripts sometimes had to contain redundant content, e.g. descriptions of common methods and algorithms, to make them more readily accessible for a general readership. Keeping these redundancies between the chapters would have made reading the thesis unnecessarily tedious, so they have been largely removed and substituted by appropriate cross references. Therefore, some chapters appear more concise than the manuscripts they are based on.

The PLOS ONE article from 2019 that forms the basis for Chapter 3, 4 and 6 had unfortunately not been published correctly. During the final typesetting stage, the PLOS ONE staff obviously slipped and crippled two important formulas. These errors were "fixed" at PLOS ONE by issuing a separate correction statement while keeping the original article with the wrong formulas despite our vigorous demand to depublish and replace the erroneous version.[2] In this thesis, the correct formulas are presented as originally intended. Several chapters also feature other minor improvements over the original manuscript versions, like fixed typing errors or amended stylistic aberrations. Apart from that, the content of this thesis closely resembles the content of the article manuscripts and, importantly, preserves improvements that were made to initial article drafts during peer-review processes.

[2]It goes without saying that PLOS ONE had fundamental reasons of overriding importance for handling correction issues this way, and I respect them for staying true to their policy.

Chapter 2.

Theoretical background and motivation

This chapter motivates the subsequently presented research, explains the required theoretical background, and defines the research goals pursued in this thesis.

The prediction of human behavior is a highly promising but challenging objective, as Subrahmanian and Kumar (2017) acknowledged. Predictions about a specific person's memory lapses and action errors with respect to given tasks could not only help human teachers or coaches to focus their instruction on each trainee's weak points but also be fed into a wide spectrum of technical assistance systems to support user-specific adaptation. The following chapters report on investigations that belong to an overarching research line investigating how anticipatory assistance systems can facilitate cognitive aspects of human activities and human-machine interaction. Prototypical application scenarios for this are in-car driver information systems and AR smart glasses overlaying the real world with virtual content. In such contexts, giving excessive step-by-step assistance for a task by constantly placing vast amounts of visual information within the users' field of view could be annoying and distracting at best. In worse cases, it may even turn out dangerous when subsystems of human cognition with limited capacity, such as those related to attention, are required to deal with too many different (or complex) sources of input in parallel. In this context, ***attention*** refers to cognitive systems that select some information (e.g. visual or auditory stimuli) from a larger set at so-called *serial bottlenecks*, i.e. points where

parallel processing of all available information is not possible (Anderson, 2020). Technical systems that always assist each and every step of an activity may also lead to a high degree of dependence on the system and impede learning processes when users resort to mindlessly following a system's instructions. For example, Maguire, Woollett, and Spiers (2006) had shown that London taxi drivers' acquisition of navigation knowledge increased their hippocampal[1] volume, whereas ten years later McKinlay (2016) warned that over-reliance on automatic wayfinding like GPS satellite-navigation systems erodes our natural abilities. Therefore, the amount of information presented to users should be restricted to the required minimum. This generally conforms with established principles from disciplines such as human-centred design (ISO 9241-110; ISO 14915), human-computer interaction (Shneiderman et al., 2016), ergonomics (ISO 15005), and usability engineering (Nielsen, 2005). To this end, it must be determined in which situations assistance is actually required. This may be the case when users are either unsure about what to do, or when they are about to do something wrong. In perilous or time-critical task sequences, these situations should obviously be anticipated beforehand to mitigate possible damage. In non-critical activities, feasible predictions could contribute to smoother task execution, better user experience and better performance rather than waiting for human errors to occur and trying to correct them afterwards. Technical systems that incorporate such an "anticipatory module", combined with effective assistance features, can induce a new level of learning processes. The subsequent chapters of Part I shall propose and evaluate new computational approaches for generating such predictions on the basis of ***structural-dimensional analysis of mental representations*** (SDA-M; see Chapter 3 and Schack, 2012) related to specified tasks.

2.1 Mental representations

A structured cognitive basis that integrates person, environment, and task information is necessary to plan and act in a goal-oriented way (see e.g. Nitsch,

[1]The hippocampus is a lateral brain structure that plays an essential role in the persistent storage of new memories (Anderson, 2020).

2004; Schack & Hackfort, 2007). In the middle of the 19th century, classical ideas in psychology (see James, 1890; Lotze, 1852) led to the *ideomotor* approach, which distinguished the important role of a cognitive equivalent of actions in memory:

> *"[...] there is no a priori difficulty in believing that Ideas [sic] may become the sources of muscular movement [...]"*
>
> (Carpenter, 1852, p. 152)

In a similar vein, Prinz (1997) proposed a framework for action control and action planning, coined the *common coding approach*, which contended that perceived events and planned actions share a common representational domain. From a cognitive-perceptual perspective, ***mental representations*** can be considered as the cognitive basis to organize, store in memory, and execute complex motor actions and movements in terms of their anticipated sensory effects (Schack & Mechsner, 2006). In recent times, different lines of research in cognitive psychology, philosophy, cognitive robotics and other disciplines refer to the central role of mental representations in action organization with different definitions and perspectives (e.g. Maycock et al., 2010; Rosenbaum, Cohen, Jax, Weiss, & Van Der Wel, 2007; Schack & Ritter, 2009, 2013). For the purpose of this thesis, it seems useful to refer to mental representations as a functional structure that integrates both perceptual and cognitive features to achieve context-specific action goals (Schack & Ritter, 2009).

A seminal theoretical framework for movement control by Bernstein (1967) described the multiple ways to reach a movement goal as a degrees-of-freedom problem. Bernstein developed a task-dependent evolutionary-originated multi-level model of movement control. Cognitive aggregations and chunking reduce the planning cost and facilitate action and movement control (Anderson, 1982; Chase & Simon, 1973). From this point of view, mental representations overcome the complexity of redundant environments to control complex movements and action sequences, leading to task-related order formation. The idea of a hierarchical cognitive architecture has since been investigated using diverse approaches (e.g. Anderson, 1983; Hoffmann, 2003; Jeannerod, 2004; Rosenbaum, 2009).

Level	Main function	Subfunction	Means
IV: Mental control	Regulation	Volitional initiation control strategies	Symbols; strategies
III: Mental representation	Representation	Effect-oriented adjustment	Basic action concepts (BACs)
II: Sensorimotor representation	Representation	Spatial-temporal adjustment	Perceptual effect representations
I: Sensorimotor control	Regulation	Automatization	Functional systems; basic reflexes

Table 2.1: **Levels of action organization** (modified from Schack, 2004, p. 408).

2.1.1 The cognitive action architecture approach (CAA-A)

A suitable model for the research presented in this thesis was proposed by Schack (2004). The model of the cognitive architecture of action uses a goal-oriented approach of regulatory levels and representational levels that are functionally autonomous (Schack & Ritter, 2013). This so-called *cognitive action architecture approach* (CAA-A; see Table 2.1) differentiates between two regulatory levels: *Sensorimotor control* (level I), which initiates lower level processes like automatized movements and reflexes, and *mental control* (level IV), which initiates volitional and control strategies (IV) (see also Frank, Land, & Schack, 2016; Land, Volchenkov, Bläsing, & Schack, 2013). The representational levels of *sensorimotor representation* (II) and *mental representation* (III) build the cognitive information basis. Perceptual effects and their spatial-temporal features are stored on the sensorimotor representation level (II), whereas the cognitive units of complex actions, the so-called ***basic action concepts*** (BACs), are located on the level of mental representation (III). Analogously to the notion of concepts of objects (Schack & Mechsner, 2006), BACs can be seen as the building blocks of motor memory that connect movement goals and recognizable perceptual effects (Schack & Ritter, 2009, 2013).

A number of studies have investigated the essential role of BACs in long-term memory in manual actions (Stöckel, Hughes, & Schack, 2012), sports actions (Bläsing, Tenenbaum, & Schack, 2009; Schack & Mechsner, 2006), sports tactics (Lex, Essig, Knoblauch, & Schack, 2015), and rehabilitation

(Braun et al., 2007; Jacksteit et al., 2017). The results characteristically show that mental representations of people with a high level of competence and expertise tend to form well-integrated hierarchical structures that are in line with the biomechanical structure and other demands of the task. In contrast, the mental representations of novices, young children or stroke patients reveal less hierarchically organized cognitive structures.

These findings are supported by experiments from Land et al. (2013) regarding modularity in motor control (d'Avella, Giese, Ivanenko, Schack, & Flash, 2015), which indicated a clear structural relationship between mental representation and the kinematic structure of movement. Furthermore, current projects and investigation on job-related knowledge have been conducted (Schack & Ritter, 2013; Seegelke & Schack, 2016; Vogel & Schack, 2016). It is assumed that, as for tactical knowledge and complex actions, the structures of working tasks in occupational rehabilitation are similarly stored in memory (Lex et al., 2015) and change over the course of learning (Frank, Land, & Schack, 2013). The mental representation structure of such tasks can be investigated by applying the SDA-M method, but as discussed later, the traditional standard procedures of this method require substantial manual effort and expertise on the part of the investigator (see Chapter 3). Subsequent chapters will explore how SDA-M can be further automatized based on approaches from complementary *computational* cognitive architectures.

2.2 Computational cognitive architectures

A major branch of research in modern cognitive science deals with the creation of unifying frameworks called ***computational cognitive architectures***, which specify the *"structure of the brain at a level of abstraction that explains how it achieves the function of the mind"* (Anderson, 2009, p. 7) and can be simulated by a corresponding computer program. Researchers and authors commonly omit the prefix "computational" and refer to these frameworks simply as "cognitive architectures". However, for the purpose of this thesis it seems important to establish a notational distinction between frameworks that are defined in such a way that they can be directly simulated by computer programs (i.e. "computational cognitive architectures") and archi-

tectures that are not directly computer-simulatable, like CAA-A, because they are defined on a verbal–conceptual level of abstraction that defies easy and exhaustive translation into program code. The general nature and purpose of computational cognitive architectures can then be defined as follows:

> *"[A computational] cognitive architecture is the overall, essential structure and process of a domain-generic computational cognitive model, used for a broad, multiple-level, multiple-domain analysis of cognition and behavior. In particular, it deals with componential processes of cognition in a structurally and mechanistically well defined way."*
>
> (Sun, 2004, p. 342)

Interestingly, Sun, as well as Anderson, differentiates between "cognitive architectures", which describe general structures and processes, and specific (cognitive) "models", which usually result from a domain- or task-dependent parametrization and instantiation of an architecture they are based on. However, cognitive architectures are obviously also models themselves, i.e. human-made abstractions from the actual or presumed constitution of the mind.

Although many computational cognitive architectures strive to incorporate, or at least consider, as much as possible of what is known about the human mind and cognition at the time of their creation, it is important to note that every architecture is based on a distinct set of assumptions. These assumptions may be based on scientific data, philosophical thoughts and arguments, as well as "computationally inspired" working hypotheses (Sun, 2004). For example, the ***Soar*** architecture (Laird, 2012) uses a modest set of building blocks, e.g. different types of learning mechanisms (*reinforcement, chunking, semantic learning, episodic learning*) and memories (*procedural, semantic, episodic*) to approximate "human-level intelligence", and it can be run on virtual or embodied agents (Trafton et al., 2013). When declarative memories need to be retrieved, Soar always simply selects the one with highest activation. If multiple possible "actions" (represented by so-called "production rules") match the current contents of working memory, they are

fired in parallel. With respect to these aspects Trafton et al. (2013, p. 31) suggested that *"Soar is concerned more with high-level functionality than with low-level cognitive fidelity, which makes it less suited to predicting people's errors and limitations."*

A recent review of the past 40 years of research on cognitive architectures reported on 84 different architectures out of which 49 were still actively developed, but only a small number of these implement rather extensive sets of capabilities to pursue the goal of achieving "artificial general intelligence" (Kotseruba & Tsotsos, 2020). As Sun (2004) and others noted, the arguably most successful among these computational cognitive architectures so far is Anderson's *"Adaptive Control of Thought – Rational"* (ACT-R; see e.g. Anderson, 2009).

2.2.1 Adaptive control of thought – rational (ACT-R)

According to the ACT-R theory (Anderson et al., 2004; Anderson & Lebiere, 1998), human behavior is predominately controlled by a central *production rule system*, which is neurophysiologically associated to a part of the brain known as the *basal ganglia*. Functionally it is related to procedural knowledge as it represents possible actions as production rules, i.e. "IF–THEN" rules. These rules take current goals, sensory inputs and chunks from declarative memory into account by matching the left side of rules ("IF") with the contents of buffers associated with the respective subsystems called "modules". This includes modules responsible for declarative memory and its retrieval, intentional goal and control functions, visual object recognition and location tracking, and a motor system (see also Figure 2.1[2]). A part of the basal ganglia, the *striatum*, is supposed to perform pattern matching functions on the contents of these modules' buffers. The right sides of rules ("THEN") describe possible actions. Overall, this symbolic level describes which actions are in principle applicable in a given situation.

ACT-R then draws on a subsymbolic layer to decide which of the applicable actions shall be executed. Another portion of the basal ganglia, the *pallidum*, selects among the applicable actions by stopping to inhibit the cor-

[2]Note that Figure 2.1 shows module boxes very roughly at those brain areas that are mentioned first below each module's name, but other regions elsewhere are also involved.

responding cells in the associated *thalamus*, which serves as a *"relay station for motor and sensory information from lower areas"* (Anderson, 2020, p. 18) and "executes" actions through projections to the respective representations in the *cortex* (Anderson et al., 2004). ACT-R's subsymbolic layer is a lower-level (computational) abstraction related to these neural processes. A very similar mechanism is used for selecting one of several chunks from declarative memory when a specific type of long-term memory content is required. These subsymbolic processes are modeled using the same basic mathematical approach: The ACT-R mechanisms for selecting production rules and for selecting memory chunks both use the Boltzmann distribution as a ***softmax*** rule for conflict resolution when more than one rule or chunk is applicable (Anderson & Betz, 2001).[3] This results in the following type of equation for calculating the probability that a chunk i is retrieved, or a production rule i is selected:

$$P_i = \frac{e^{X_i/s}}{\sum_j e^{X_j/s}} \tag{2.1}$$

In this equation, X_i is substituted either by ACT-R's formula for *chunk activation* (A_i) or the formula for *production rule utility* (U_i). Based on rational analyses and theoretical arguments (see e.g. Anderson & Schooler, 1991), chunk activation is supposed to reflect the log-odds that an applicable chunk will be matched in the present context (Anderson, 1993; Anderson & Lebiere, 1998; Lebière, Anderson, & Reder, 1994). In ACT-R models, it is calculated as

$$A_i = B_i + \sum_j W_j S_{ji} \tag{2.2}$$

where B_i represents the general "base-level" activation of chunk i, each W_j is an attentional weighting of a memory element j that belongs to the current goal chunk, and S_{ji} represents the strength of association between the element j and chunk i.

Production rule utility (U_i) can be understood either as the rationally estimated worth of the production rule, or as a trade-off function between ex-

[3]To be precise, the softmax equation for chunk choice that is presented here is actually a closed-form approximation of the behavior of ACT-R models, which generate predictions by Monte Carlo simulations (Anderson & Betz, 2001).

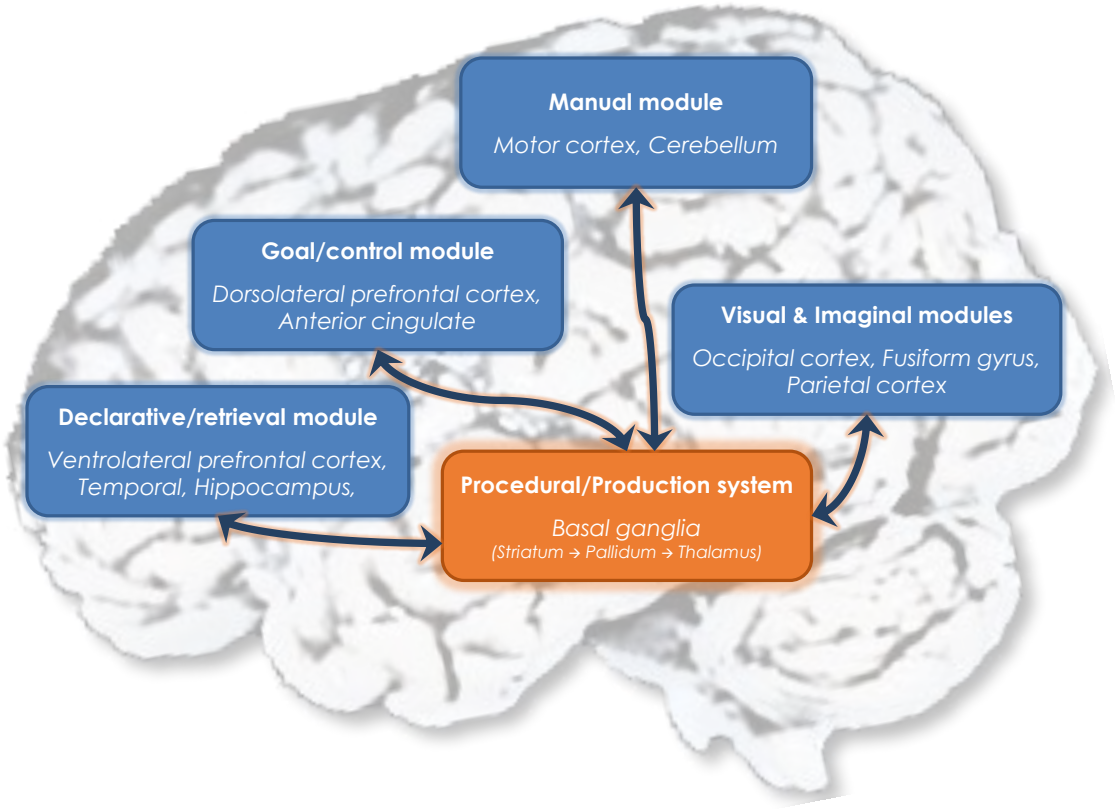

Figure 2.1: **Modular organization of the ACT-R architecture** and associated brain regions (adapted and consolidated from Anderson et al., 2004, 2008).

pected costs and the value of the goal (Anderson & Lebiere, 1998):

$$U_i = P_i G - C_i \tag{2.3}$$

where P_i is the learned likelihood that production rule i would achieve the current goal, G represents the goal's value, and C is the estimated cost (e.g. the required time). Additionally, ACT-R makes use of a *production strength* value S_i, which is supposed to measure the log-odds that rule i will fire, and is adapted to speed up its firing depending on frequency of use.

The softmax equation (2.1) incorporates a constant $s > 0$ that reflects noise in the context of chunk activation and is typically set at 0.4 in ACT-R (Anderson et al., 2004). This noise value s plays an analogous role to the "temperature" value in Boltzmann machines or simulated annealing (Hinton, 2007): The higher s, the less preference is given to items with higher values.

Based on simulations and mathematical analyses of asymptotic properties, Belavkin (2001, p. 52) suggested that the values G and s in ACT-R may represent psychological *arousal*, i.e. *"the activation or the "energy" of a cognitive process"*, whereas the ratio G/s could be understood as an indicator of the "confidence level" of a problem solver.

2.3 Human-centered system development and ethics

A substantial body of current research is concerned with finding proper ways of educating and sensitizing engineers to ethics (e.g. Bairaktarova & Woodcock, 2017; Cheruvalath, 2017; Gelfand, 2016; Miñano, Uruburu, Moreno-Romero, & Pérez-López, 2016; Murphy & Gardoni, 2017; VanDeGrift, Dillon, & Camp, 2017). However, less extensive guidance has been offered regarding approaches to systematic handling of ethical issues during actual development processes for systems based on information and communication technology (ICT). While traditional software engineering process models like the *waterfall model* were divided into discrete, sequential phases (Royce, 1987), proponents of ***agile methodologies*** like Scrum's inventor Ken Schwaber have rejected this:

> *"The stated, accepted philosophy for systems development is that the development process is a well understood approach that can be planned, estimated, and successfully completed. This has proven incorrect in practice."*
>
> (Schwaber, 1997, p. 117)

In order to cope with uncertainty and limited plannability, and react flexibly to changing requirements, agile approaches do not consider system features, architectures and components as static and fixed throughout development. Instead, planning of development tasks is limited to short timeframes and continually readjusted (Beck, 2000). Since results of usability tests and correspondingly required design changes are hardly predictable beforehand, especially when they directly involve prospective users as participants, agile development methodologies generally accord very well with the requirements

of human-centered design processes. A large number of well-defined process models have been developed for combining agile development approaches like *Extreme Programming* (Beck, 2000) or *Scrum* (Schwaber, 1997) with user-centred design methods (e.g. Holzinger, Errath, Searle, Thurnher, & Slany, 2005; Lee, McCrickard, & Stevens, 2009; Memmel, Gundelsweiler, & Reiterer, 2007; Obendorf & Finck, 2008; Singh, 2008). These user-centred design methodologies strongly emphasize the importance of usability as a product characteristic but disregard any ethical aspects that do not happen to coincide with specific user requirements regarding effective, efficient and satisfying system usage. As of yet, few specific guidelines are offered on how to assess and handle ethical issues during the day-to-day work in agile development processes that are characterized by transient requirement definitions and limited overall predictability.

Arguably the most well-known approach that aims at tackling some of these issues is the *Value-Sensitive Design* (VSD) methodology (Friedman, Kahn, & Borning, 2008), which has been applied in more or less structured ways in many projects (e.g. Friedman et al., 2008; Royakkers & Steen, 2016; Umbrello & De Bellis, 2018; van den Hoven, Vermaas, & van de Poel, 2015). VSD is presented as a tripartite methodology comprising three types of value-related "investigations" (conceptual, empirical, and technical), which "overlap and intertwine so that boundaries between them are blurred" (Davis & Nathan, 2015, p. 32). Publications on VSD (e.g. Friedman et al., 2008) claim that stakeholders and benefits/harms for these must be identified, mapped onto corresponding values, and should be explicitly related to relevant design trade-offs. However, only recently (see also Manders-Huits, 2011; Reijers et al., 2017; Yetim, 2011) a suitably comprehensive overview was published about which methods and tools could be applied to these ends (Friedman, Hendry, & Borning, 2017). Overall, the selection and systematic integration of appropriate methods in agile development processes still remains an underspecified aspect in the "official" VSD literature by Batya Friedman and colleagues. An elaborate, well-defined methodology that connects VSD approaches with IT system design processes has been proposed by Spiekermann (2015). While her *ethical system design lifecycle* (E-SDLC) fits classical, plan-driven development processes particularly well, Spiekermann (2015,

p. 164) claimed that *"agile software development can [also] be used in ethical system design. The only thing that needs to be fulfilled is that earlier system design phases get the requirements and architecture right up front."* This may not be feasible in many projects, since development teams often choose agile approaches when they expect frequent requirements changes and commonly re-factor the code to adjust the architecture correspondingly (see also Beck, 2000).

Overall, this methodological lack constitutes a pressing issue, because an absence of explicit ethical considerations may lead to suboptimal adoption of new technologies such as intelligent assistive systems (Ienca, Wangmo, Jotterand, Kressig, & Elger, 2017), e.g. smart glasses for cognitive assistance.

2.4 Research questions

Based on the theoretical background and limitations of existing approaches, the following high-level research questions can be defined for this work:

- **RQ1**: Can the procedures for analyzing task-related mental representation structures based on SDA-M (see Chapter 3) be further automatized with algorithmic approaches (e.g. from computational cognitive architectures)?
- **RQ2**: Do these "algorithmic SDA-M" analyses conform to the gold standard of "traditional SDA-M" that involves human expert assessments?
- **RQ3**: How well can algorithmic SDA-M analyses predict human errors related to different kinds of practical applications (e.g. manual assembly tasks or movement sequences)?
- **RQ4**: Are algorithmic SDA-M analyses sensitive to changes in memory formation (e.g. caused by learning processes)?
- **RQ5**: Are algorithmic SDA-M analyses applicable irrespective of skill levels, i.e. equally suitable for experts and laypersons in a particular domain?

- **RQ6**: How well can algorithmic SDA-M analyses assess people's formal expertise and overall performance in an activity compared to traditional SDA-M-based measures?
- **RQ7**: Which methodological procedures should be used to take ethical issues and other system stakeholder requirements properly into consideration when developing cognitive assistance systems (e.g. smart glasses incorporating an algorithmic SDA-M component)?

Table 2.2 below indicates which subsequent chapters address each of these research questions. The respective chapters contain only implicit but obvious references to the research questions, although the RQ numbers are not stated. An explicit overview about the obtained "answers" to each of these research questions will finally be given in the thesis' general discussion (Chapter 11).

Part	Chapter	Addressed research questions
-	1	-
-	2	-
I	3	-
I	4	RQ1
I	5	RQ1
I	6	RQ2
I	7	RQ3, RQ4, RQ5
I	8	RQ3, RQ6
II	9	RQ7
II	10	RQ7
-	11	-

Table 2.2: **Research questions addressed by thesis chapters.**

PART I

COMPUTATIONAL ANALYSES OF MENTAL REPRESENTATION STRUCTURES

Chapter 3.

The SDA-M method

This chapter is based on parts of:

Strenge, B., Vogel, L., & Schack, T. (2019).
Computational assessment of long-term memory structures from SDA-M related to action sequences. *PLOS ONE, 14*(2). Public Library of Science. doi:10.1371/journal.pone.02124140

Abstract

Structural-dimensional analysis of mental representations (SDA-M) is an established method for retrieving human memory structures related to specific activities. For this purpose, SDA-M involves a semi-automatized survey of users (the "split procedure"), which yields data about users' associations between action representations in long-term memory. This data about associations has commonly been clustered and visualized by SDA-M software in the form of dendrograms that can be used by human experts as a tool to (manually) assess users' individual expertise and identify potential issues with respect to predefined action sequences. This chapter explains methodical and computational details of the SDA-M method and motivates further automation of that process.

3.1 Standard procedures

As mentioned earlier, the SDA-M method can be used to analyze human memory structures related to a given set of items (e.g. actions). Well-integrated cognitive networks lead to more structured decisions in the SDA-M split procedure. The method then provides psychometric data that can be analyzed on an individual and on a group level. To this end, the standard procedure of SDA-M comprises a preliminary task analysis and up to four steps (Lander, 1991; Schack, 2012), which are outlined in the following.

3.1.1 Task analysis

In a preparatory step, it is generally important to understand the activity and characterize its task-adequate functional organization, usually in collaboration with novices, practitioners or professionals of different levels of expertise, and coaches. When SDA-M is used to analyze a specific activity, the

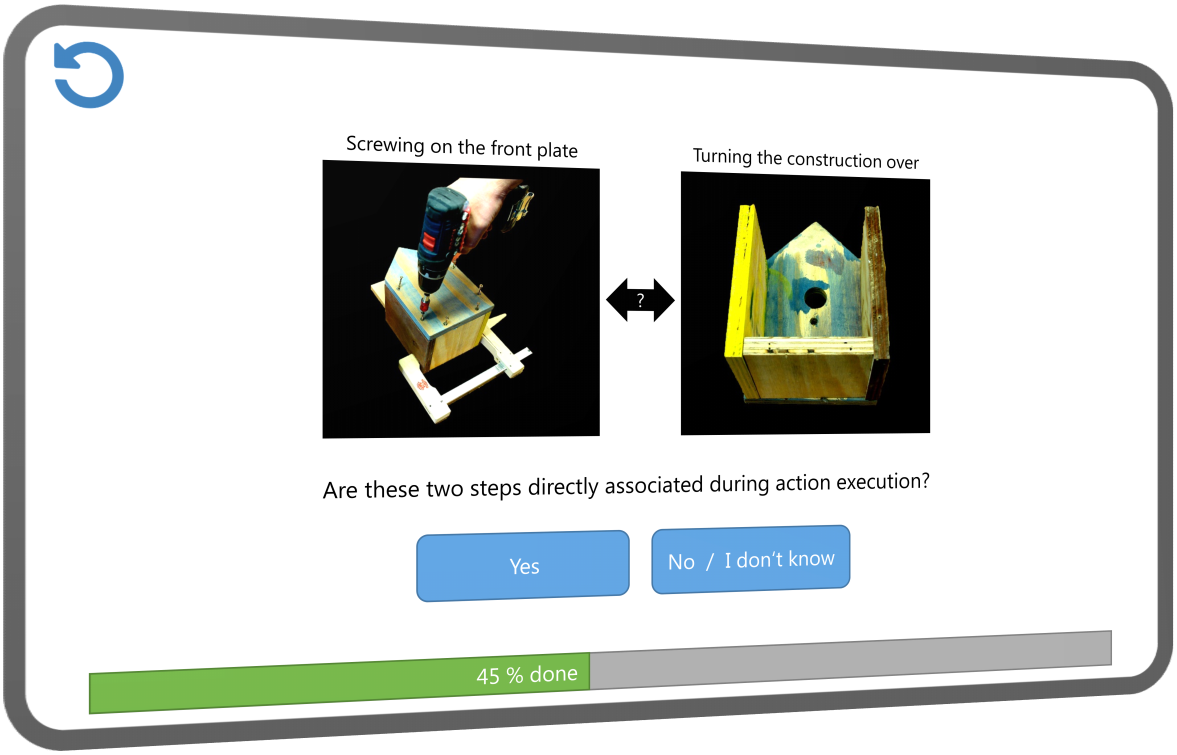

Figure 3.1: **QSplit SDA-M tool UI concept.** This illustration of the QSplit SDA-M tool's user interface concept for performing split procedures on mobile devices shows two exemplary action representations related to the activity 'building a birdhouse'.

activity is first split into basic actions which are indicated by textual descriptions, pictures or illustrations, short video clips, or a combination of those means. This can be done by researchers with the help of a functional movement analysis (Hossner, Schiebl, & Göhner, 2015) and in collaboration with domain experts (e.g. coaches) to compile a "plausible and workable set" of BACs (Schack, 2012).

3.1.2 Step 1: Split procedure and distance scaling

The split procedure technique is based on the selection and presentation of a set of BACs that comprises a valid and necessary subset of the larger set of concepts for the activity or domain from which they stem. These action items (BACs) are then shown to study participants or users, usually on a computer screen. Figure 3.1 shows the split procedure user interface concept for a mobile touch-friendly version of the QSplit SDA-M software. Actions are chosen in random order as reference objects or "targets" and then, one after another, all other actions are compared to the current target in random order. The user must decide for each pair of actions (a_i, a_j) whether these are directly associated during task execution or not. The decisions made in the context of each target result in a particular decision tree, i.e. in the end the number of decision trees is equal to the total number of actions n. For each target action a_i let A_i be the subset of actions the user considered as "associated" to the target, including a_i itself. The split procedure then creates an X matrix consisting of n row vectors $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{in})$. The following equation (3.1) describes the exact values that are assigned when the common "fast" or "1-level" split procedure is used:

$$x_{ij} = \begin{cases} |A_i| + 1 & \text{if } i = j \\ 1 & \text{if } a_j \in A_i, i \neq j \\ |A_i| - n & \text{otherwise} \end{cases} \tag{3.1}$$

Multiple splitting steps may be performed for each reference action in order to yield a more fine-grained distance measure. However, most contemporary applications of SDA-M, including this study, are restricted to only one splitting step for each reference action in order to reduce the required time

and effort for participants. Lander and Lange (1992) argued that a metrically defined measure of distance from a reference object (target) to any other can be obtained by standardizing the respective x values to z-scores, thus establishing a "Z matrix" containing one such (row) vector of z-scores for each action. The SDA-M software then creates matrices containing the correlations ("R matrix") and Euclidean distances ("D matrix") between all rows of the Z matrix. The distance values in the D matrix (or, equivalently, the correlation values in the R matrix) contain all information to completely define an individual's representational structure (Tscherepanow et al., 2011). The subsequent steps of SDA-M are therefore functions of these matrices.

3.1.3 Step 2: Hierarchical clustering and visualization

The distances calculated in the first step are now used as the metric for hierarchical agglomerative average-linkage clustering. The results are visualized by a dendrogram (as shown on the right side of Figure 6.1) to facilitate human assessment of the mental representation structure. For many SDA-M applications this is the last necessary analysis step.

3.1.4 Step 3: Extraction of feature dimensions

This step aims to uncover the latent criteria or feature dimensions that seem to have guided subjects' decisions during the split procedure. To this end, the R matrix is subjected to factor analysis with a special cluster-oriented rotation procedure (Lander, 1991; Schack, 2012).

3.1.5 Step 4: Analysis of interindividual differences

Pairs of individual or subgroup-specific clustering results (representing mental representation structures) can be analyzed to determine their invariance or degree of similarity. For this purpose, Lander and Lange (1992) and Schack (2012) proposed the structural invariance measure λ. This requires that step 2 (but not necessarily step 3) has previously been finished. Let the sets S_a and S_b represent the outcomes of SDA-M's hierarchical agglomerative average-linkage clustering for participant a and participant b, which contain the clusters $C_i \in S_a$ and $C_j \in S_b$ of BACs. The invariance of the mental

representation structures of participants a and b is then defined as follows:

$$\lambda_{a,b} := \sqrt{\frac{\min(|S_a|, |S_b|)}{\max(|S_a|, |S_b|)} \cdot \frac{\sum_{i=1}^{|S_a|} \sum_{j=1}^{|S_b|} |C_i \cap C_j|}{\sum_{i=1}^{|S_a|} \sum_{j=1}^{|S_b|} \sqrt{|C_i| \cdot |C_j|}}}; \lambda_{a,b} \in [0, 1] \quad (3.2)$$

More recently, the Adjusted Rand Index (ARI) gained popularity among SDA-M researchers for measuring the similarity of two participants' mental representation structures (see e.g. Frank, Land, Popp, & Schack, 2014; Frank et al., 2013, 2016; Jeraj, Musculus, & Lobinger, 2017; Kim, Frank, & Schack, 2017; Land, Frank, & Schack, 2014; Meier, Frank, Gröben, & Schack, 2020). The ARI is bounded above by a maximum of 1 and takes on negative values (with no well-defined lower bound) when similarity falls below the expected value from random clustering with the same number of clusters and elements in each (Hubert & Arabie, 1985).

3.2 Usage in individual cognitive assessment and coaching

Numerous previous studies have indicated that educated psychologists, sports scientists, mathematicians, and domain experts could use visualizations of mental representation structures from SDA-M (i.e. the dendrograms from step 2) to detect individual issues regarding action execution and derive helpful advice for performance optimization (e.g. Heinen & Schack, 2004; Heinen & Schwaiger, 2002; Heinen, Schwaiger, & Schack, 2002; Schack, 2004; Schack & Hackfort, 2007). The SDA-M method enables addressing individual needs by taking the essential information about the underlying cognitive-perceptual action system into account (Schack & Hackfort, 2007). For example, mental representations related to gymnastics skills were retrieved from novices and experts. Individual mistakes in carrying out the movement were analyzed based on SDA-M data. It was reported that individual interventions based on those mental representations accelerated and optimized the learning process and brought novices' mental representation structures closer to those of experts (Frank et al., 2013, 2016; Heinen et al., 2002). The SDA-M method has been applied to numerous activities in manual action, sports, dancing and rehabilitation (Bläsing et al., 2009; Frank et al., 2016; Schack, 2004; Schack & Ritter, 2013; Weigelt, Ahlmeyer, Lex, &

Schack, 2011) to investigate expertise-dependent memory structures and develop related individualized training strategies (Schack, 2020; Schack, Essig, Frank, & Koester, 2014; Schack & Hackfort, 2007).

This line of research provided evidence that SDA-M data visualized as dendrograms can be interpreted by appropriately trained human specialists (psychologists, mathematicians etc.) to identify deficits in memory structures. On account of this, the next chapter describes how SDA-M data can be *automatically* interpreted by a technical system to trigger corresponding assistance when needed.

Chapter 4.

Advanced algorithmic approaches

This chapter is based on parts of:

Strenge, B., Vogel, L., & Schack, T. (2019).
Computational assessment of long-term memory structures from SDA-M related to action sequences. *PLOS ONE, 14*(2). Public Library of Science. doi:10.1371/journal.pone.02124140

Abstract

This chapter presents new algorithmic approaches for automatizing the process of assessing task-related memory structures based on SDA-M data to predict probable errors in action sequences. Two alternative algorithmic approaches to human error prediction based on SDA-M data have been developed. These shall be called *Analysis of Most Probable Actions* (AMPA), and *Correct Action Selection Probability Analysis* (CASPA), respectively. Formal analyses of the approaches outline their commonalities and differences on a theoretical level.

4.1 Assumptions and prerequisites

Both algorithmic approaches (AMPA and CASPA) require as input

- a predefined list of all correct action sequences (related to an activity), and
- valid SDA-M data for a specific person X (related to an activity).

The output of the algorithms then indicates when (i.e. after which actions) person X may require assistance while performing the activity.

Furthermore, both algorithms require the overarching activity or task to be represented in SDA-M through a set of n subtasks or actions *("BACs")* satisfying the following criteria:

- *Atomicity*: Each action is self-contained insofar as it is assumed to be executable by each person without issues. If this was not the case, it must be divided further into feasible sub-actions before performing the SDA-M split procedure. The resulting BACs can be understood as problem-solving operators available to users.
- *Sequential discreteness*: Actions do not overlap in time. All correct sequences of actions can be formed by strictly ordering a subset of all actions.
- *Non-recurrence*: Each action appears at most once in each correct action sequence. (Note: In practical applications this restriction can often be worked around by adding sequential information to descriptions of identical actions in the SDA-M split procedure, e.g. *"Pressing the yellow button for the first time"* and *"Pressing the yellow button for the second time"*.)
- *Completeness*: The total set of actions considered during the SDA-M split procedure comprises all actions that can be executed while performing the activity.
- *Context-independence*: Environmental and contextual factors not explicitly incorporated into action descriptions do not influence behavior.

- *Currentness*: The SDA-M data for a given person is valid in the sense that his or her task-related memory structure has not changed since the SDA-M split procedure was performed.

In practical applications these theoretical assumptions may not hold to full extent, hence decreasing the achievable accuracy of predictions but not necessarily rendering the results unusable. For example, the assumption of *completeness* will commonly be violated to some degree by focusing, for pragmatic reasons, on a set of *probable* task-related actions instead of all *possible* actions. This is inevitable because the SDA-M split procedure (the "manual" part of the method) has a time complexity of $\Theta(n^2)$, i.e. the time for performing it grows quadratically as a function of the number of actions. According to practical experience this usually limits the number of incorporable actions to approximately 10-15 (depending on the time required for each decision), because subjects are rarely willing to perform split procedures lasting much longer than quarter of an hour. In a similar vein Tscherepanow et al. (2011) stated that the number of actions "should not be chosen higher than 20. Otherwise, the decisions made regarding the similarity of stimuli may become inconsistent". The requirement of sequential discreteness must be accounted for when determining the actions ("BACs"). Furthermore, participants should be disposed to ideally associate each action exactly with what they believe to be the immediate preceding and subsequent actions with respect to correct sequences. To this end, the current version of the QSplit SDA-M software incorporates an introductory video (in German) that instructs participants to state whether the displayed actions are executed immediately before or after another during task execution. Note that many but not all previous applications of SDA-M complied with the requirement of sequential discreteness (Braun et al., 2007; Jacksteit et al., 2017; Schack, Essig, et al., 2014).

4.2 Algorithm I: Analysis of Most Probable Actions (AMPA)

The first step of SDA-M involves calculating a measure of distances between any two of the analyzed items (e.g. objects or actions) in a person's long-

term memory. Algorithm I determines whether there is a *correct* immediate follow-up action which has lowest distance among *all* actions (or second-lowest distance in the case that the second-last action has lowest distance to the last executed action), which equates to the strongest association between these actions. We call this a *"Correct Most-Probable Action"* (CMPA), being aware that there may be more than one CMPA in any situation. If there are no CMPAs in a current situation then it is probable that the person will either choose an (incorrect) action with stronger association or not know how to proceed, i.e. assistance is required. The concept of assuming that exactly those chunks which have the highest activation ($\widehat{=}$ lowest distance) are always chosen is very straightforward and may seem highly simplified given the noisy nature of human behavior. Nonetheless it constitutes a promising heuristic; e.g. it has successfully been used as a basic assumption for a computational cognitive model of instance-based learning (Said, Engelhart, Kirches, Körkel, & Holt, 2016), as well as by the Soar cognitive architecture (Laird, 2012).

To formalize this approach, let $n \in \mathbb{N}$ be the total number of actions related to the considered task and $A = \{a_1, ..., a_n\}$ the set of all these actions. Let $S \subset A$ be the set of all actions a specific person has already executed in a given situation, including action $a_i \in S$ as the second-most recent one and $a_j \in S$ being the most recent one. Let $C_S \subseteq A \setminus S$ be the set of all correct immediate follow-up actions in this situation, and D_{a_x,a_y} the distance between any two actions a_x and action a_y in the person's memory (as calculated by SDA-M; see section 3.1.2). Then the value of competent(S) indicates whether in this situation, after action a_j, the person is assumed to know what to do next on their own:

$$\text{competent}(S) := \begin{cases} 1 \quad \text{if } \exists a_c \in C_S: \ \forall (x \in \mathbb{N} \mid x \leq n \wedge x \neq i,j): \\ \qquad\qquad\qquad\qquad D_{a_j,a_c} \leq D_{a_j,a_x} \\ 0 \quad \text{otherwise} \end{cases} \tag{4.1}$$

In this formula, action a_c is a CMPA. Note that it is not required that there is a correct action with strictly smaller distance than all other actions, but only

that it is *among* those actions closest to the most recent one. Action a_j itself as well as its immediate predecessor a_i are hereby disregarded (in contrast to less recent actions from set S). Since SDA-M's pairwise distance values are undirected, it would be neither unexpected nor detrimental to task execution if a_i had lower distance to a_j than all correct follow-up actions, but it seems rather improbable that a_i would be repeated after a_j. With respect to these aspects, AMPA is an *optimistic* heuristic.

As an example, assume that exactly these two action sequences are correct for some task:

$$\begin{aligned} \text{seq}_1 &:= (a_1, a_2, a_3, a_4, a_7) \\ &\text{and} \\ \text{seq}_2 &:= (a_1, a_2, a_3, a_5, a_6). \end{aligned} \tag{4.2}$$

Now assume that a person has already executed the actions (a_1, a_2, a_3) with S being the set of this tuple's elements. If, among all actions, the most recently executed action a_3 has lowest distance to its predecessor a_2, then action a_3 must have *second-lowest* distance to either action $a_4 \in C_S$ or action $a_5 \in C_S$ for the person to be considered "competent" in this situation. If not, action a_3 must have *lowest* distance to a_4 or a_5. Otherwise the person would be deemed unable to determine a correct follow-up action. For example, if a_6 is closest to a_3 in memory, i.e. $\arg\min_{a_x}(D_{a_3,a_x}) = a_6$, the person would probably try to execute action a_6 after action a_3, which would be wrong.

4.3 Algorithm II: Correct Action Selection Probability Analysis (CASPA)

In contrast to the AMPA algorithm, CASPA does not only output a plain binary assessment of competence in a given situation but a continuous measure of probability. This allows for a much more fine-grained assessment of mental representation structures as well as task-, user- and context-specific thresholds for when to provide assistance. For this purpose, CASPA inherits concepts used by the ACT-R cognitive architecture (Anderson et al., 2004; Anderson & Lebiere, 1998). As discussed more extensively in Section 2.2.1, the ACT-R theory assumes that human behavior is predominately controlled

by a central production rule system, which is neurophysiologically associated to the basal ganglia. Functionally it is related to procedural knowledge as it represents possible actions as production rules, i.e. "IF–THEN" rules. These rules take current goals, sensory inputs and chunks from declarative memory into account by matching the left side of rules ("IF") with the contents of buffers associated with the respective subsystems called "modules" (see Figure 2.1). The right sides of rules ("THEN") describe possible actions. Overall this symbolic level describes which actions are in principle applicable in a given situation. ACT-R then draws on an additional subsymbolic layer to decide which of the applicable actions shall be executed. This subsymbolic layer is a lower-level abstraction related to neural processes. A very similar mechanism is used for selecting one of several chunks from declarative memory when a specific type of long-term memory content is required. Therefore it does not matter for our purposes whether the actions of a specific task covered by an SDA-M procedure are (in terms of ACT-R) more related to contents of declarative memory or to executive functions associated with the production rule system. In fact, the distinct behavior of the subsymbolic levels of these two processes is modeled using the same basic mathematical approach: The ACT-R mechanisms for selecting production rules and for selecting memory chunks both use the Boltzmann distribution as a "softmax rule" for conflict resolution when more than one rule or chunk is applicable (Anderson & Betz, 2001). As we will show now, this approach can be adapted to estimate the probability of a specific person choosing a correct action in a given situation based on SDA-M data.

4.3.1 Calculations

Let $A = \{a_1, ..., a_n\}$ be the set of all actions related to the considered activity, and $S \subset A$ the set of all actions the person has already executed in a given situation, including action $a_i \in S$ as the second-most recent and $a_j \in S$ as the most recent one. Let $C_S \subseteq A \setminus S$ be the set of all correct immediate follow-up actions in this situation, and $I_S \subseteq A \setminus C_S$ be all actions which are *applicable but incorrect* in the given situation with respect to successful task

execution. Then the probability that the person will know what to do after action a_j is estimated as follows:

$$P_S = \sum\nolimits_{a_c \in C_S} \frac{e^{\rho(a_j,a_c)/s}}{\sum_{a_x \in (C_S \cup I_S) \setminus \{a_i,a_j\}} e^{\rho(a_j,a_x)/s}} \tag{4.3}$$

This calculation incorporates a constant $s > 0$ that reflects noise and for our application is set at 0.4, which is a typical value concerning chunk activation in ACT-R (Anderson et al., 2004). This noise value s plays an analogous role to the "temperature" value in Boltzmann machines or simulated annealing (Hinton, 2007): The higher s, the less preference is given to actions with higher activation.

Equation (4.3) further requires a measure $\rho(a_x, a_y)$ representing the strength of association between actions a_x and a_y in users' memory or, in this context equivalently, the activation level of an action a_y after action a_x has been executed. Lander proposed such a measure, called π, as part of the original SDA method (Lander, 1991), the predecessor of SDA-M:

$$\pi(a_x, a_y) = \exp(-\frac{D_{a_x,a_y}}{D_{krit}}) = 1/\exp(\frac{\sqrt{1 - r_{a_x,a_y}}}{\sqrt{1 - r_{krit}}}), 0 < \pi(a_x, a_y) \leq 1 \tag{4.4}$$

A drawback of this formula is that the value of π depends on the "incidental correlation value" r_{krit} as defined by Schack (2012), which in turn depends on an arbitrarily chosen significance level α as well as the total number of actions. Furthermore, uncorrelated and even negatively correlated actions (i.e. $r_{a_x,a_y} \leq 0$) still show association strength values $\pi(a_x, a_y) > 0$, no matter which r_{krit} value is determined (see Figure 4.1). Overall the slope of the function leads to insufficient discrimination between negatively or weakly correlated items and moderately correlated ones. To mitigate these issues, an alternative calculation based on the ACT-R formulas for production strength and chunk activation can be used. These formulas reflect the log-odds that an applicable chunk will be matched in the present context, or that an instantiation of a production rule will fire (Anderson, 1993; Anderson & Lebiere, 1998; Lebière et al., 1994).

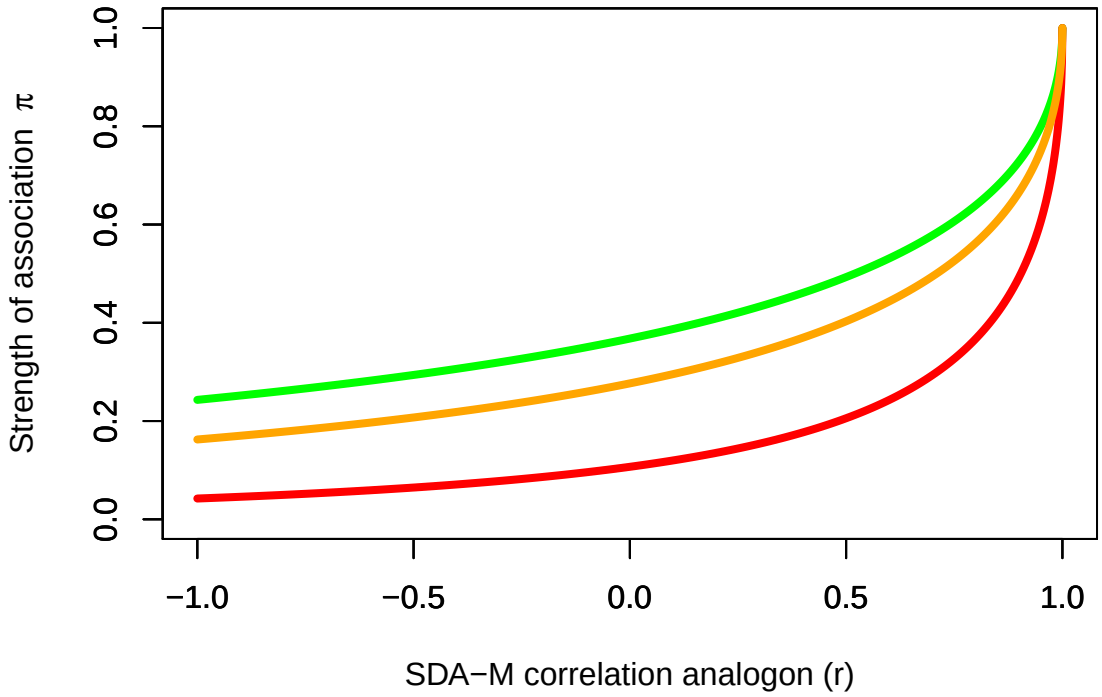

Figure 4.1: **Relation of Lander's association strength measure π to the SDA-M correlation analogon r.** Green: $r_{krit} = 0$. Orange: $r_{krit} = 0.39$. Red: $r_{krit} = 0.8$.

For our purposes, the SDA-M correlation analogon r is used analogous to the respective probability values in ACT-R (see Figure 4.2):

$$\text{Activation } \rho(a_x, a_y) := \log(\frac{r_{a_x,a_y}}{1 - r_{a_x,a_y}}) \tag{4.5}$$

Finally, Equation (4.3) is adjusted to take those actions into consideration which are positively correlated to (i.e. associated with) the most recent action a_j, such that CASPA regards these as applicable to the given situation in terms of the ACT-R theory:

$$P_S = \sum\nolimits_{a_c \in C_S,\ r_{a_j,a_c}>0} \frac{e^{\rho(a_j,a_c)/s}}{\sum_{a_x \in A \setminus \{a_i,a_j\},\ r_{a_j,a_x}>0} e^{\rho(a_j,a_x)/s}} \tag{4.6}$$

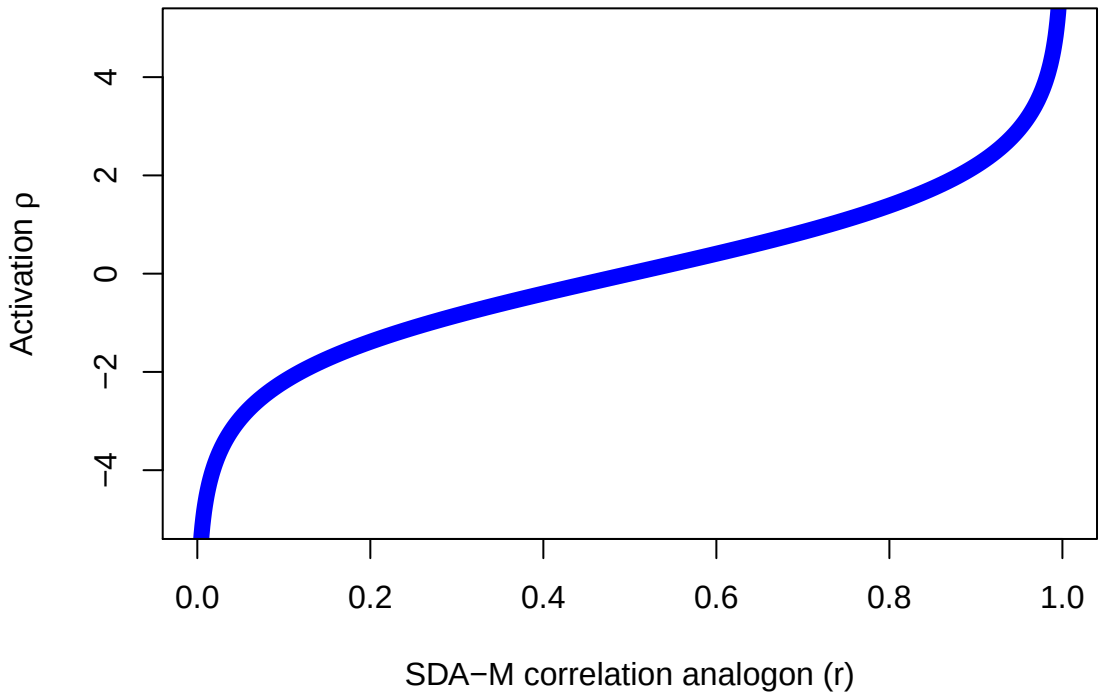

Figure 4.2: **Relation of activation measure ρ to the SDA-M correlation analogon r.**

4.3.2 Default vs. informed threshold

In order to decide whether assistance should be given, arbitrary thresholds for P_S can be used. The most natural approach a priori would be to choose a probability threshold of 0.5, i.e. whether it is supposedly more likely that assistance is needed or dispensable. However, it may be beneficial to determine an *informed threshold* setting using empirical data if available. To this end, a sufficiently large number of SDA-M data sets for the respective task must be available so that the average probability estimated for different situations by CASPA sufficiently converges. The threshold is then set to that average estimated value of P_S. The assumed benefit of this is the compensation of possible systematic biases of P_S as determined by CASPA. Such systematic biases may occur when the SDA-M split procedure is slightly easier or harder for subjects to perform than the real task due to artifacts of modeling the real

actions in the form of visual and/or textual representations. This approach also takes another potential issue into account: Theoretically, the value of P_S should to some degree be dependent on the total number of actions considered during the SDA-M split procedure. Assuming purely random decisions on part of the subject, it holds that the more actions are included in the split procedure, the lower the expected value of P_S. In practice subjects may induce such bias through random tie-breaking in case of doubt as well. Concerning the final binary decision regarding competence or feedback, an empirically informed threshold may mitigate these issues. In the following the binarized output of CASPA using the default threshold (0.5) will be referred to as CASPA$_d$ while an informed threshold will be denoted as CASPA$_i$.

4.4 Relations between the algorithms

In the following, let the sets C_S and I_S contain only "applicable" actions which are positively correlated with (i.e. associated to) the most recent action a_j, i.e. exactly those considered by CASPA (see Equation 4.6). In some special cases the output of CASPA is identical to that of AMPA:

- All applicable follow-up actions are correct:
  $|C_S| \geq 1 \wedge I = \emptyset$
  $\Rightarrow$ output(AMPA) = output(CASPA) = 1.
- There is no correct applicable action:
  $C_S = \emptyset$
  $\Rightarrow$ output(AMPA) = output(CASPA) = 0.
- The noise value is set at $s \to 0$ and among the applicable actions with maximum activation is no incorrect action but at least one correct:
  $(\forall a_i \in I_S : \rho(a_j, a_i) < \max_{a_x} \rho(a_j, a_x)) \wedge (\exists a_c \in C_S : \max_{a_x} \rho(a_j, a_x) = \rho(a_j, a_c))$
  $\Rightarrow$ output(AMPA) = output(CASPA) = 1

- The noise value is set at $s \to 0$ and among the actions with maximum activation is no correct one:
  $\nexists a_c \in C_S : \rho(a_j, a_c) = \max_{a_x} \rho(a_j, a_x)$
  $\Rightarrow$ output(AMPA) = output(CASPA) = 0

It should be remarked that since the regular noise value for CASPA is constant at $s = 0.4$ the relations depending on zero noise are merely theoretical statements. Generally the following relations hold:

- If there is a correct applicable action with maximum activation:
  $\exists a_c \in C_S : \rho(a_j, a_c) = \max_{a_x} \rho(a_j, a_x)$
  $\Rightarrow 0 <$ output(CASPA) $\leq$ output(AMPA) = 1.

- If there are correct applicable actions, but none of these has maximum activation:
  $C_S \neq \emptyset \wedge \forall a_c \in C_S : \rho(a_j, a_c) < \max_{a_x} \rho(a_j, a_x)$
  $\Rightarrow 0 =$ output(AMPA) $<$ output(CASPA) < 1.

Because of these relations it is not possible to tell in general which algorithm is “more optimistic” or “more pessimistic”, or to derive the output of one algorithm from the other algorithm’s output.

“Nothing is as practical as a good theory”

– Kurt Lewin (1945)

Chapter 5.

Theoretical remarks on CASPA

This original chapter complements the line of argument from the previous chapter. It explains more details of the rationale behind CASPA's cognitive modeling and algorithm design decisions based on theoretical arguments.

The statement that CASPA's calculations use the SDA-M correlation analogon r analogous to the respective probability values for chunk matching in ACT-R (see Section 4.3.1) may warrant some further theoretical explanations. It may also not be directly evident why CASPA's final equation discards negative correlations and only takes those actions into consideration that are positively correlated to (i.e. associated with) the most recent action. Both of these algorithm design decisions can be motivated based on the two following theorems.

Theorem 1

Lander (1991) suggested that a mental representation structure calculated based on an SDA split procedure and resulting correlation matrix was "both structurally and metrically" invariant to the mental representation structure calculated based on a similarity rating scale, as long as both were obtained from the same subjects. In both cases the same internally-represented structure regarding existing relations and feature assignments was activated, i.e. both methods are supposed to map the same structure (Lander, 1991). To this end, similarity judgments $S_{ij} \in [a, b]; a, b > 0$ with a representing "com-

pletely different" and b representing "very similar" had to be linearly mapped to correlation values ranging from 0 ("completely different") to 1 ("very similar") using the following equation[1]:

$$r_{ij} = \frac{(S_{ij} - a)}{b - a}$$

Note that this mapping disregards negative correlation values. Similarly, the CASPA algorithm also disregards negative and zero correlations. In explaining the factor analysis part of his method, Lander (1991) also argues that SDA correlation values specify the degree of association or the closeness of connection between conceptual elements in memory due to their weighting of corresponding features.

Theorem 2

It seems reasonable from a connectionist point of view to quantify the likelihood that an active chunk activates another chunk through spreading activation in terms of how many elements are commonly associated to both of these chunks. Adopting ACT-R terminology (cf. Anderson et al., 2004), an active chunk (representing a basic action concept that is currently used for action execution) can likewise be understood as a (partial) representation of the "context" for selecting the next action in an action sequence. When no further information about the strengths of associations or base-level activations is available, and assuming fixed chunk sizes for simplicity, the proportion of common elements suggests itself as an estimate of the probability of matching chunks. A related symmetric similarity measure for binary vectors is known as the Kulczynski index (see e.g. Batagelj & Bren, 1995; Zakani, Arhid, Bouksim, Gadi, & Aboulfatah, 2016), which is based on the conditional probability that a specific feature occurs in one item, given that it occurs in the other.

Figure 5.1 shows a hypothetical example in which chunk $C1$ spreads activation to its neighbors $C2$ and $C3$. Considering the number of common

[1] Lander obviously slipped and confused the constants a and b in his equation. Therefore, a corrected version is presented here, as it was undoubtedly intended.

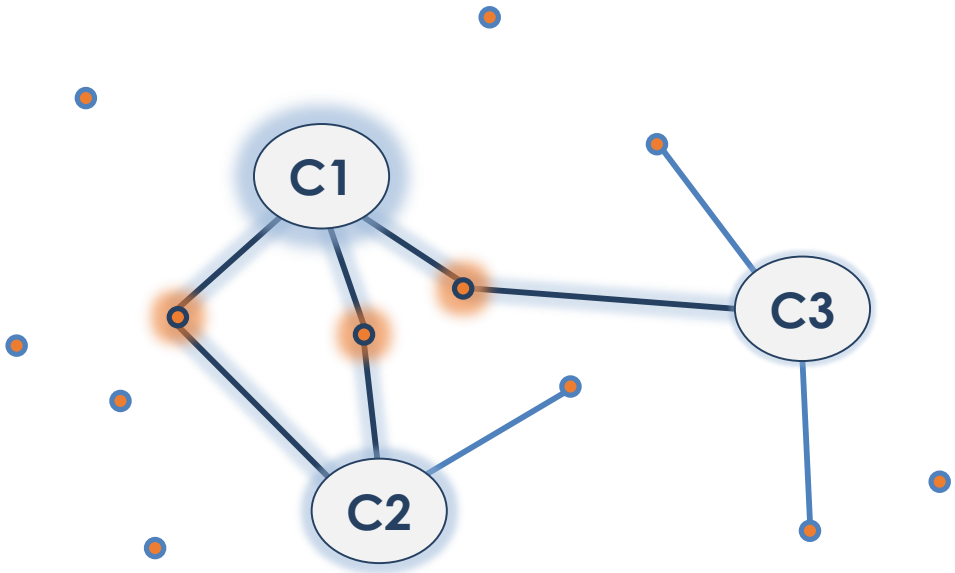

Figure 5.1: **Hypothetical constellation of chunk activations via matching elements.** Chunk $C1$ spreads activation to chunks $C2$ and $C3$ through common associations.

associations in absence of further information suggests that the expected likelihood of $C2$ becoming activated by $C1$ should be twice as high as $C3$'s. This also appears neurally plausible to some degree: If one assumes that a neuron receives a random (unknown) amount of electrochemical excitation within an arbitrary fixed interval $R := (0, x]$ from each of its n associated neurons and has a random (unknown) threshold $t \in (0, y], y \geq \frac{xn}{2}$, then the probability that the expected accumulated input exceeds the threshold t causing the neuron to fire grows proportionally to the number n of associated elements.[2]

Based on these observations the CASPA algorithm assigns a matching probability of 1 to feature vectors representing perfectly matching associations, which correspond to chunks that are identical in terms of their associations. Conveniently, the SDA-M procedure also implicitly adds a self-association with maximum value to each action.

As mentioned before, the SDA(-M) method was originally designed to not only gather dichotomous decisions from "fast" 1-level split procedures but

[2]This statement disregards inhibitory input and assumes sampling from uniform distributions within the defined intervals, i.e. the sum of input activation values follows a linearly transformed Irwin–Hall distribution with an expected value of $\frac{1}{2}xn$.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	.495	1								
3	.325	.66	1							
4	.232	.49	.747	1						
5	.183	.386	.589	.797	1					
6	.139	.314	.484	.66	.83	1				
7	.11	.262	.409	.555	.707	.853	1			
8	.093	.221	.345	.474	.608	.737	.871	1		
9	.068	.183	.298	.413	.534	.654	.77	.879	1	
10	.052	.156	.265	.364	.468	.583	.682	.786	.896	1

Table 5.1: **SDA-M correlation values between vectors from 1-level splitting** for different number of associations or chunk sizes (rows; 2 to 10) and number of common associations (columns; 1 to 10) assuming a total of $n = 200$ actions. The values obviously approximate the ratio between the number of common associations and chunk size.

also permits multi-level splitting to yield fine-grained metric data. For the sake of consistency between 1-level and multi-level splitting, the initial value assignments to $\mathbf{x}$ vectors resulting from the dichotomous choices in 1-level split procedures are not represented as binary but integer vectors (see Equation 3.1). The Kulczynski index is therefore not directly applicable to SDA-M data. However, the values of SDA-M's correlation analogon between vectors with a common total number of associations converge against Kulczynski similarity index values for 1-level splits with large numbers of actions as shown exemplarily in Table 5.1.

Under these assumptions and conditions an alternative interpretation of the SDA-M correlation analogon as an approximation of Kulczynski similarity index values for the corresponding binary vectors is warranted.[3] In conjunction with the previous argument about the relation between the number of associations and likelihood of generating action potentials, this substantiates the interpretation of (positive) SDA-M correlation values as approximate probability estimates in the context of chunk matching mechanisms in CASPA.

[3]Note that negative correlations are disregarded or mapped to zero.

Chapter 6.

Comparison with human experts' assessments

This chapter is based on parts of:

Strenge, B., Vogel, L., & Schack, T. (2019).
Computational assessment of long-term memory structures from SDA-M related to action sequences. *PLOS ONE, 14*(2). Public Library of Science. doi:10.1371/journal.pone.02124140

Abstract

This chapter reports on a first evaluation study, which compared automatized assessments by the AMPA and CASPA algorithms to predictions made by human scholars based on visualizations of SDA-M data. The different algorithms' outputs matched human experts' manual assessments in 84% to 86% of the test cases.

As mentioned before, previous studies have demonstrated that human experts (scholars) could use specific visualizations of mental representation structures based on SDA-M data to detect individual issues regarding action execution and derive helpful advice or training concepts for performance optimization. This substantiates the assumption that feasible algorithms would achieve the same if they interpreted SDA-M data in a way that conforms to interpretation by humans. The study described in this chapter investigated to which degree the different computational approaches from Chapter 4 satisfy this criterion.

6.1 Data base

In order to establish a suitable test set of SDA-M data as a data pool for further analyses, we cooperated with a local diaconal non-profit foundation working with people with various mental disorders. In a first step, relevant working tasks related to preparing, opening and cleaning a kiosk at the foundation were identified by observing the operational procedure. These tasks had been used by the foundation as part of an educational program for people with mental disorders for several years. In the second step we interviewed two coaches to detect the underlying working structure. In the third step the amount of working steps was reduced by integrating similar and related steps. In the next step the set of concepts was tested in a pilot study. At the end, the set of working tasks was adjusted and retested. Afterwards, these items were applied to the SDA-M software. A total of 27 trainees with mental disorders, comprising depression, schizophrenia, substance use disorders, autism spectrum disorders, attention deficit hyperactivity disorder, anxiety and mood disorders, used the software to judge whether a pair of actions belongs together during their work in the kiosk. All participants gave informed consent in written form. Their capacity to do so was ensured by asking our contacts at the foundation (trained professionals in coaching people with disabilities) to exclude all trainees for whom this might be questionable. In the SDA-M splitting procedure, a total of 15 different actions were covered, which could be divided into four independent activities:

Kiosk preparation:

- Refill cutlery cart
- Fill in coffee beans and cocoa
- Refill fridge with drinks
- Put plate for rolls in place
- Allocate cart for dirty dishes

Kiosk customer service:

1. Welcome the customer and take the order
2. Prepare coffee and cocoa
3. Serve drinks and food
4. Take the money

Kiosk wrap-up:

- Wash the dishes and start the dishwasher
- Clean the glass pane of the refrigerator
- Clean the coffee machine
- Wipe the surfaces

Laundry:

1. Wash the laundry
2. Hang up and iron the laundry

The actions related to customer service and laundry naturally have to be executed in a sequential order (as indicated by the numbering above). Therefore, ideally these actions should pairwise be strongly associated in long-term memory structures to represent the correct sequence. Actions related to the kiosk preparation and wrap-up activities can be executed in arbitrary order (indicated by bullet points). Generally, actions related to different activities should ideally not be associated to each other in long-term memory.

The trainees were familiar with these actions and activities to differing degrees, because they were trained in these tasks at the diaconal organization for different lengths of time (between a few weeks and several months). In line with previous studies, e.g. on actions in judo (Weigelt et al., 2011), windsurfing (Schack & Hackfort, 2007), soccer (Schack & Bar-Eli, 2007) or manual actions in humans and robots (Schack & Ritter, 2009, 2013), we assume that potential problems and deficits in action execution are reflected in the mental structure of the tasks. Thus, unrelated or wrongly related actions on the cognitive level are expected to lead to decreased real-life performance, e.g. forgetting of the next relevant action or executing a wrong task. For example, a trainee might start cleaning a table instead of serving a customer who is waiting in line.

6.2 Retrieval of experts' manual assessments

The data pool was then used to compare assessment by human SDA-M experts with algorithmic interpretation. To this end, an assessment task consisting of 80 different hypothetical situations related to the kiosk-servicing activities listed above was created. Each of these "situations" was specified by

- SDA-M data visualization of a random subject, and
- a fictitious sequence of actions this subject was said to have executed up to now.

The fictitious action sequences had been created by selecting representative subsets from the set of all correct sequences and applying a random cut-off length to each sequence. By definition, a "correct sequence" was a sequence

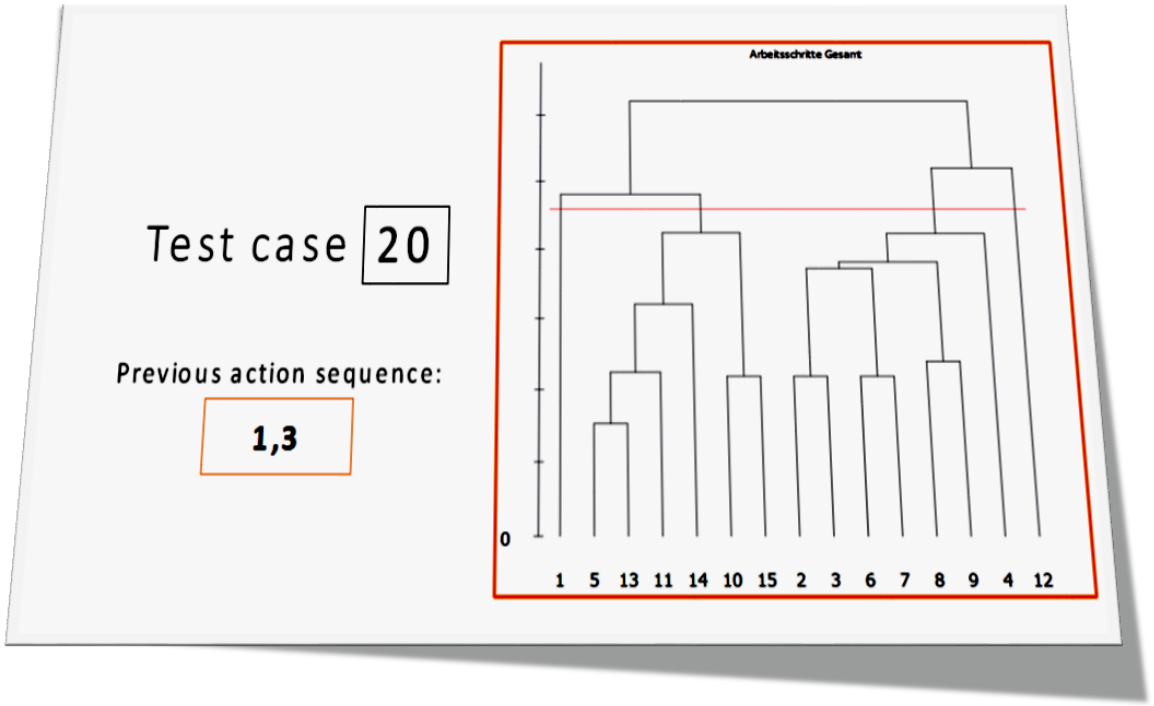

Figure 6.1: **Example of a test case representing a fictitious "situation" for assessment.** The right side shows a dendrogram visualizing a subject's mental representation structure for the kiosk service activities. In this example some actions from the "kiosk preparation" activity (IDs 2, 3 and 4) are clustered with actions from the "customer service" activity (IDs 6, 7, 8 and 9), indicating a corresponding relation in memory.

containing actions from only one of the four activities and, where applicable, in correct temporal order. The first half of the final set of sequences was initially chosen randomly. The resulting set was then manually revised to mitigate a bias towards sequences from the larger, unordered activities caused by the disproportionate number of permutations of actions in these activities. The second half was determined by randomly selecting from a set of sequences that was priorly adjusted by adding duplicates of some sequences to compensate for over-/underrepresentation of activities. These "situations" or test cases were then presented (as shown in Figure 6.1), one after another, to a group of $N = 12$ human scholars, along with a general overview of all correct sequences for each of the four kiosk-service-related activities. The participating scholars were experts with extensive education regarding the SDA-M method and personally experienced in using it for scientific purposes before, but they were blind with respect to the algorithmic analyses that were

investigated in this study. Each scholar had to assess independently for each situation, based on the given SDA-M data visualizations, whether or not the respective subject would more likely need assistance or be able to determine a correct follow-up action in the given situation. The same test cases were also fed into the AMPA and CASPA algorithms. As both experts and algorithms pursued the same goal (predicting human errors), their results could then be compared as described in the next section.

6.3 Data analysis and results

The assessments by each of the human experts have been translated into binary vectors with value 1 representing the assessment *"the subject in this scenario is probably able to determine a correct follow-up action in the given situation on their own, i.e. assistance is not required"*, and value 0 representing the opposite case. The assessments from 11 out of 12 human SDA-M experts correlated positively with the group average, whereas those from one expert correlated negatively. Presumably this was due to misunderstandings regarding the assessment task. Therefore this expert's ratings were excluded from further analyses. The remaining assessments served as the ground truth for comparison with the respective results from the AMPA and CASPA algorithms.

As CASPA delivers estimated probability values $P_S \in [0, 1]$, a direct comparison was possible with the portion of experts $P_E \in [0, 1]$ who supposed in each test case that the respective subjects were competent. A positive correlation of $r = .62$ was found, which, considering that the mean correlation (determined using Fisher z-transformation) of each individual expert's assessments to the average assessments of the remaining experts was almost identical ($\overline{r} = .59$), indicates an adequate fit between manual and algorithmic assessments.

In order to evaluate the (binary) output from AMPA and the influence of different thresholds for CASPA, several common metrics for the evaluation of binary classifiers have been employed. For this purpose the Median value of the experts' assessments was used for each test case. Due to an odd number of experts ($N = 11$) this equals majority decision. CASPA's continuous P_S

values were converted into binary decisions as described in section 4.3.2, i.e. using either the default threshold of 0.5 ("CASPA$_d$") or an informed threshold of $\overline{P_S} = 0.2396$ ("CASPA$_i$"), where $\overline{P_S}$ was the average of all probability values output by CASPA for all 80 test cases from the study. In the following, let N_{ab} with $a, b \in \{0, 1\}$ be the total number of situations where the human experts' assessment equals a and an algorithm's prediction equals b. The simple matching coefficient (SMC) for binary vectors yields the percentage of cases where human and algorithmic assessments came to the same results, thus representing the accuracy of matching the human experts' assessments regarding expected action errors:

$$\text{Accuracy} = \text{SMC} = \frac{N_{00} + N_{11}}{N_{00} + N_{01} + N_{10} + N_{11}} \tag{6.1}$$

One-tailed binomial tests with $H_1 : P(\text{success}) > P(\text{failure})$ were performed for each algorithmic approach to determine whether the degree of match between human and algorithmic assessments, i.e. the accuracy, is significantly above chance level. Each matching pair of assessments counted as a successful Bernoulli trial and each deviating pair as a failure. Correlations between the respective vectors of binary decisions were calculated and tested for significance as well.

Sensitivity, specificity, and positive/negative predictive values can be defined analogously to accuracy (Equation 6.2 to 6.5). In this context a "true positive" denotes cases of both algorithm and human experts suspecting that assistance was required because the subject's mental representation structure is not suitable (N_{00}).

$$\text{Sensitivity} = \frac{N_{00}}{N_{00} + N_{01}} \tag{6.2}$$

$$\text{Specificity} = \frac{N_{11}}{N_{11} + N_{10}} \tag{6.3}$$

$$\text{Positive predictive value (PPV)} = \frac{N_{00}}{N_{00} + N_{10}} \tag{6.4}$$

$$\text{Negative predictive value (NPV)} = \frac{N_{11}}{N_{11} + N_{01}} \tag{6.5}$$

In addition to these classic metrics the *balanced accuracy* should be considered, because this measure safeguards against biased classifiers taking advantage of an imbalanced test set (Brodersen, Ong, Stephan, & Buhmann, 2010). If an algorithm performs equally well in terms of sensitivity and specificity, its balanced accuracy reduces to the conventional accuracy.

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{N_{11}}{N_{11}+N_{10}} + \frac{N_{00}}{N_{00}+N_{01}}\right) \tag{6.6}$$

Table 6.1 shows how AMPA, CASPA$_d$ (threshold = 0.5), and CASPA$_i$ (threshold = 0.2396) performed with respect to these metrics. The conventional accuracy values (Figure 6.2) were close to the balanced accuracy values with all algorithms ranging between 0.78 and 0.86 for these metrics. Binomial tests showed that with all three algorithm variants the match with human experts' assessments was highly significant above chance level. Differences between the algorithms were marginal, though CASPA$_i$ generally tended to score slightly better than AMPA.

6.4 Discussion

The analysis of mental representation structures using the SDA-M method is a well-established approach for gaining insight into the degree of individual expertise related to various activities, ranging from basic grasping actions to complex system interactions (see e.g. Bläsing et al., 2009; Braun et al., 2007; d'Avella et al., 2015; Frank et al., 2013; Jacksteit et al., 2017; Land et al., 2013; Lex et al., 2015; Schack, 2012; Schack & Mechsner, 2006; Schack & Ritter, 2009, 2013; Seegelke & Schack, 2016; Stöckel et al., 2012). Tradition-

Algorithm	**Correlation**	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Balanced accuracy**
AMPA	0.61^{***}	0.84^{***}	0.86	0.77	0.91	0.68	0.82
CASPA$_d$	0.61^{***}	0.85^{***}	0.93	0.64	0.87	0.78	0.78
CASPA$_i$	0.67^{***}	0.86^{***}	0.88	0.82	0.93	0.72	0.85

*** $p < 0.0001$

Table 6.1: **Full results of the evaluation study.**

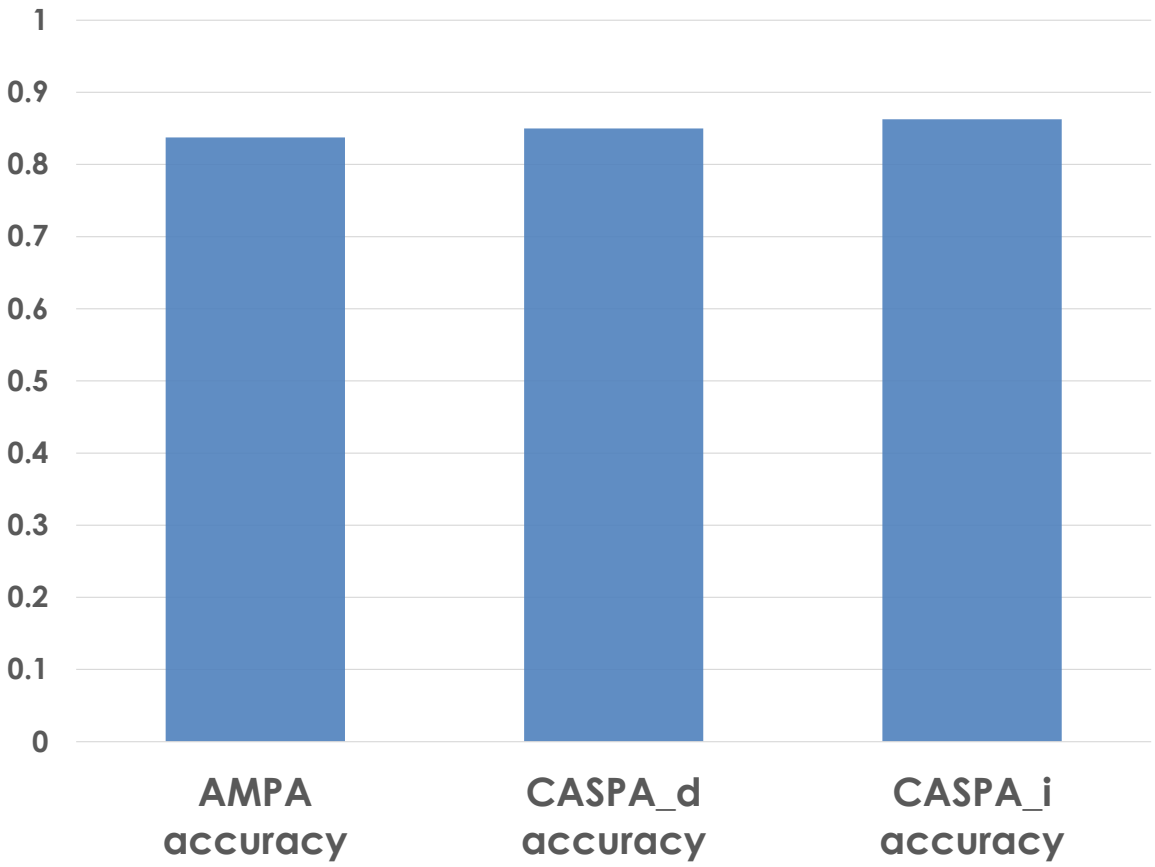

Figure 6.2: **Key results of the expert evaluation study.** The accuracy values indicate the congruence of algorithmic (AMPA, CASPA$_d$, CASPA$_i$) and human experts' assessments.

ally this information was computationally pre-processed and visualized to be interpreted by human SDA-M experts. As this requires human resources, specific training and is time-demanding, this approach is inefficient, nondeterministic and not applicable in real-time systems. Therefore we investigated different approaches to algorithmically automatize the interpretation of SDA-M data. In order to enable suitable predictions about error-prone steps during task execution, specific prerequisites must be satisfied. Most notably, the considered activity must be divisible into a limited set of sequential actions or sub-tasks which can be assumed to be executable without issues. When used as a component of a technical assistance system, this approach is most advantageous if the expected benefits from error predictions outweigh spending approximately 10-15 minutes for performing the SDA-M split procedure before system usage. This may commonly apply when executing the assisted

actions in reality is relatively time-consuming and/or when errors have severe consequences, e.g. when wrong actions are difficult to reverse. Presumably it might also help specific target groups overcoming insecurity and hesitation to tackle unfamiliar activities. In order to take learning processes into account and further reduce unneeded assistance, users may want to update the data about their mental representation structures from time to time by repeating the SDA-M split procedure.

In a first evaluation study, the proposed algorithms for SDA-M-based error prediction, AMPA, CASPA$_d$ and CASPA$_i$, showed a high degree of consistency with human experts' assessments about probable action errors based on SDA-M visualizations of subjects' mental representation structures. The percentage of matches between algorithmic and experts' assessments was significantly higher than would be expected by chance, ranging from 84% to 86%. The differences between the proposed algorithmic variants were insignificant, but the more sophisticated CASPA$_i$ algorithm scored slightly higher regarding all considered metrics than the simpler AMPA algorithm. It should be noted that the existence of some non-matching cases did not necessarily imply that the respective algorithmic predictions were wrong. On the one hand, human experts also varied from one another in their judgments regarding error predictions to some degree. On the other hand, some of the information contained in the raw data is lost when visualizing mental representation structures via dendrograms for manual interpretation. On this account the algorithmic interpretations may actually have been better than those from human experts. However, due to a lack of definitive ground truth regarding the actual mental structures of subjects from this study, this hypothesis can neither be confirmed nor rejected so far. Generally, the evaluation study reported in this chapter constitutes a proper indication of suitability of the algorithmic interpretations of SDA-M data in comparison with the traditional approach of manual assessment for a specific task. Noteworthy limitations of the study are the relatively small number of activities that were analyzed, as well as the present empiric evidence in favor of the new algorithmic approaches being restricted to a comparison with experts' assessments. Further research is mandatory to reliably assess the degree of match between predicted errors and human errors actually occurring during task execution in reality. Per-

taining to categorizations of human errors (Norman, 1981; Reason, 1990), we expect the approach to cover most (knowledge- and rule-based) mistakes, and potentially also some types of slips, e.g. due to associative activation and capture errors (excluding external event sources), loss of activation and faulty triggering. However, since many occurrences of slips are context-dependent and unreproducible, the SDA-M split procedure certainly cannot be expected to capture all (and possibly not even most) instances of slips.

“Our work becomes qualitatively
better with an intelligent assistant but
it is still our work.”

– Buchanan, Davis, and Feigenbaum (2018)

Chapter 7.

Prediction of human error in manual assembly tasks

This chapter is based on:

Strenge, B., & Schack, T. (submitted, *Scientific Reports*).
Empirical relationships between algorithmic SDA-M-based memory assessments and human errors in manual assembly tasks.

Abstract

The majority of manufacturing tasks are still performed by human workers, and this will probably continue to be the case in many industry 4.0 settings that aim at highly customized products and small lot sizes. Recent algorithmic advancements automatized the assessment of task-related mental representation structures based on SDA-M, which could enable technical systems to anticipate mistakes and assist workers during manual assembly. Two studies have empirically investigated the relations between algorithmic assessments of individual memory structures and the occurrences of human errors in different assembly tasks. Hereby theoretical assumptions of the automatized SDA-M assessment approaches were deliberately violated in realistic ways to evaluate the practical applicability of these approaches. Substantial but imperfect correspondences were found between task-related mental

representation structures and actual performances with sensitivity and specificity values ranging from .63 to .72, accompanied by prediction accuracies that were highly significant above chance level. These results are discussed in terms of practical implications and newly raised scientific questions.

7.1 Introduction

In 2020, manual assembly by human workers still plays a crucial role in many industrial areas and will likely continue to do so for many years to come. On the one hand, technical systems such as robots powered by sophisticated sensors and highly precise actuators become capable of performing more and more assembly actions autonomously. On the other hand, trends towards increased customization of products and correspondingly smaller lot sizes demand increasingly high flexibility. Humans stand heads and shoulders above machines in this regard despite impressive advancements in the field of machine learning and other artificial intelligence techniques. Unsurprisingly, the vast majority (72%) of manufacturing tasks were still performed by humans according to a recent survey report from A.T. Kearney (Hu, Akella, Kapoor, & Prager, 2018). Especially the automotive industry reportedly learned from a range of recent experiences that human workers had to be brought back to the production lines. In 2016, Markus Schaefer, head of production at Mercedes Benz, stated *"Robots can't deal with the degree of individualization and the many variants that we have today"*, so the company was *"moving away from trying to maximize automation with people taking a bigger part in industrial processes again"* (Behrmann & Rauwald, 2016). Japanese car manufacturer Toyota already initiated a similar re-introduction of manual labor a few years earlier (Trudell, Hagiwara, & Jie, 2014). On 13 April 2018, Tesla CEO Elon Musk tweeted that *"excessive automation at Tesla was a mistake"* and *"humans are underrated"*. Siemens CEO Joe Kaeser and management consulting firm Oliver Wyman therefore concordantly prognosticated that robots would not replace human workers in manufacturing anytime soon (Harbour & Scemama, 2017; Kaeser, 2017).

Working on this premise, a large body of current research is concerned with building technical systems using augmented reality setups and other

advanced technologies to assist human workers in manual assembly (e.g. Blattgerste, Renner, Strenge, & Pfeiffer, 2018; Blattgerste, Strenge, Renner, Pfeiffer, & Essig, 2017; Büttner et al., 2017; Essig, Strenge, & Schack, 2016; Evans, Miller, Pena, MacAllister, & Winer, 2017; Funk et al., 2017; Funk, Kosch, Greenwald, & Schmidt, 2015; Mura, Dini, & Failli, 2016; Renner & Pfeiffer, 2017b; Sand, Büttner, Paelke, & Röcker, 2016; Tang, Owen, Biocca, & Mou, 2003; Wang, Ong, & Nee, 2016). Ideally, such systems should show as little unneeded information as possible in order to save their human users attentional resources but provide helpful information when the worker would not know what to do, prevent them from doing something wrong, and support learning processes. These requirements make manual assembly processes interesting application scenarios for task-related human memory analyses based on the SDA-M method (see Chapter 3) and especially its recent extension by algorithmic approaches for automatized human error prediction (see Chapter 4). This chapter reports on two studies that empirically investigated the practical relations between outcomes of SDA-M-based analyses and the occurrences of human errors in manual assembly. The studies were designed in an application-oriented way and therefore entailed realistic violations of several theoretical assumptions of the SDA-M-based assessment algorithms.

As described in Chapter 3, SDA-M involves a semi-automatized survey and calculation procedure that yields user-specific data about the strength of associations between mental representations of actions in the context of a specific overarching activity. Chapter 4 described how this data can be automatically analyzed in order to assess the likelihood that a specific user would know which actions should be executed in a given situation during the activity. Three different algorithmic variants, the *Analysis of Most-Probable Actions* (AMPA) and the *Correct Action Selection Probability Analysis* (CASPA) using either a default value (CASPA$_d$) or an informed value as a decision threshold (CASPA$_i$), have been shown to be highly consistent with the conventional SDA-M approach that involves manual assessment of SDA-M data visualizations (dendrograms) and related statistical parameters by specially trained human experts. AMPA and CASPA are based on a common set of assumptions (see Section 4.1). In practical applications some of these assumptions will commonly be violated to some degree. Therefore, the

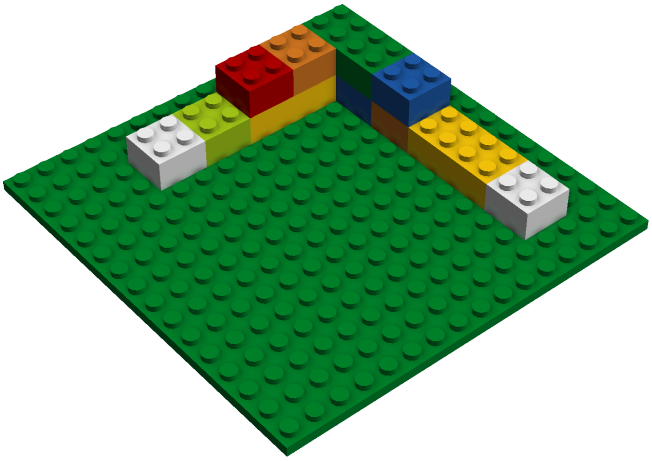

Figure 7.1: **Duplo construction** consisting of the first 12 parts of a standardized assembly task by Funk et al. (2015).

Figure 7.2: **Drawer system mockup** from Hettich.

actual accuracy of AMPA's and CASPA's predictions has been empirically investigated for two different manual assembly tasks:

1. A cheap and easily reproducible pick-and-place assembly task derived from a standardized benchmark task by Funk et al. (2015), which uses Lego Duplo bricks (see Figure 7.1), and

2. a real-world assembly task from an industrial setting, which uses parts of a drawer system mockup by company Hettich (see Figure 7.2).

This combination of tasks was chosen due to experimental feasibility, practical relevance, and so that different assumptions of AMPA and CASPA were

violated to varying degree. Table 7.1 provides a rough overview how severely each assumption was violated in the two scenarios according to a 4-point ordinal scale from *"Not violated"*, *"Not substantially violated"*, and *"Moderately violated"*, up to *"Strongly violated"*. The two scenarios were examined independently from each other as discrete studies with disjunct groups of participants, but they used similar study designs. As a foundation for both studies, participants underwent a limited phase of learning or education about the assembly proceedings. Next, their task-related mental representation structures were retrieved with SDA-M software. Finally, they were asked to execute the assembly procedures. Any errors that were made during the assembly tasks were recorded and afterwards compared to the errors that the AMPA and CASPA algorithms would have predicted based on participants' individual SDA-M data. The main differences between the two studies were related to the types of assembly actions, participants' task-related education, and the control of contextual influences.

The Duplo assembly study used an experimental design with random assignment of participants to two groups. These groups received different material during an initial learning phase to induce heterogeneity concerning their task-related knowledge. While one half of participants received a printout with completely correct instructions for assembling the designated brick construction, the other half received instructions that contained some wrong assembly steps. These erroneous instructions were meant to simulate situations in which either the available blueprints or engineering drawings for a specific construction contain minor errors, or workers engage in building a

Assumptions (see Section 4.1)	**Duplo study**	**Hettich study**
Atomicity	Not violated	Not violated
Sequential discreteness	Not violated	Not violated
Non-recurrence	Not violated	Not violated
Completeness	Strongly violated	Moderately violated
Context-independence	Not substantially violated	Strongly violated
Currentness	Not substantially violated	Not substantially violated

Table 7.1: **Degrees of violation of SDA-M algorithms' assumptions in assembly studies.**

new variant of a similar but slightly different construction they had learned to assemble in the past. The Duplo study was conducted in a quiet and controlled lab environment. In this study, a measurement of task-related memory structures with SDA-M was not only done *after* the learning phase (as in the Hettich study) but additionally also at the very beginning *before* the learning phase in order to further validate the SDA-M-based assessment procedure. Since the assembly task was unknown to participants by then, a valid assessment of their task-related memory structures at this point was expected to be characterized by low probabilities of correct action selections, and differ from the corresponding assessment after learning.

The Hettich drawer assembly study used a quasi-experimental design that distinguished between participants with either more or less extensive task-related expertise ("experts" and "laypersons"). All participants were employees of company Hettich, one of the world's leading manufacturers of furniture fittings. The experts group consisted of carpenters, joiners, and other workers with extensive task-related knowledge. The laypersons group consisted mainly of clerks, managers, and other office workers with limited professional experience in manual assembly. The study was conducted within an actual working environment at company Hettich in order to establish realistic conditions for practical assessment.

In summary, the research questions led to the following main hypothesis

- H_1: Algorithmic assessments of memory structures based on SDA-M data related to a specific assembly task correspond to subsequent outcomes of attempted action executions (success or error) in the respective assembly task,

and these supplementary hypotheses:

- H_2 (Duplo study): Algorithmic assessments of initial memory structures before learning of an unknown assembly task indicate a lack of task-related knowledge.

- H_3 (Duplo study): Algorithmic assessments of initial memory structures before learning of an unknown assembly task differ from the assessment of memory structures that are retrieved after learning.
- H_4 (Hettich study): The accuracy of algorithmic assessments of individual memory structures is independent of task-related expertise, i.e. the accuracy of prediction for "laypersons" does not differ from the accuracy of prediction for "experts".

7.2 Methods

Statement of ethical approval

Both studies have been approved by the ethics committee of Bielefeld University in written form according to the guidelines of the German Psychological Society (DGPs) and the Association of German Professional Psychologists (BDP). All participants gave informed and written consent to participate in the study.

7.2.1 Participants

Duplo study

$N_{\mathrm{D}} = 36$ individuals between 18 and 38 years with a mean age of 24.5 years ($SD = 4.3$) participated in the study. The acquisition was based on a call for participation in the form of textual announcements placed on several walls of Bielefeld University and the FH Bielefeld University of Applied Sciences. Therefore it is safe to assume that most participants were students or employees of these universities. They were either reimbursed for their time with 5 Euros in cash, or credited with one hour of experimental participation in partial fulfillment of the requirements of an eligible study program at Bielefeld University. The majority (83%) of participants were female. One half of all participants was randomly assigned to a group who received partially erroneous assembly instructions ("EI group") and the other half to a correctly instructed group ("CI group").

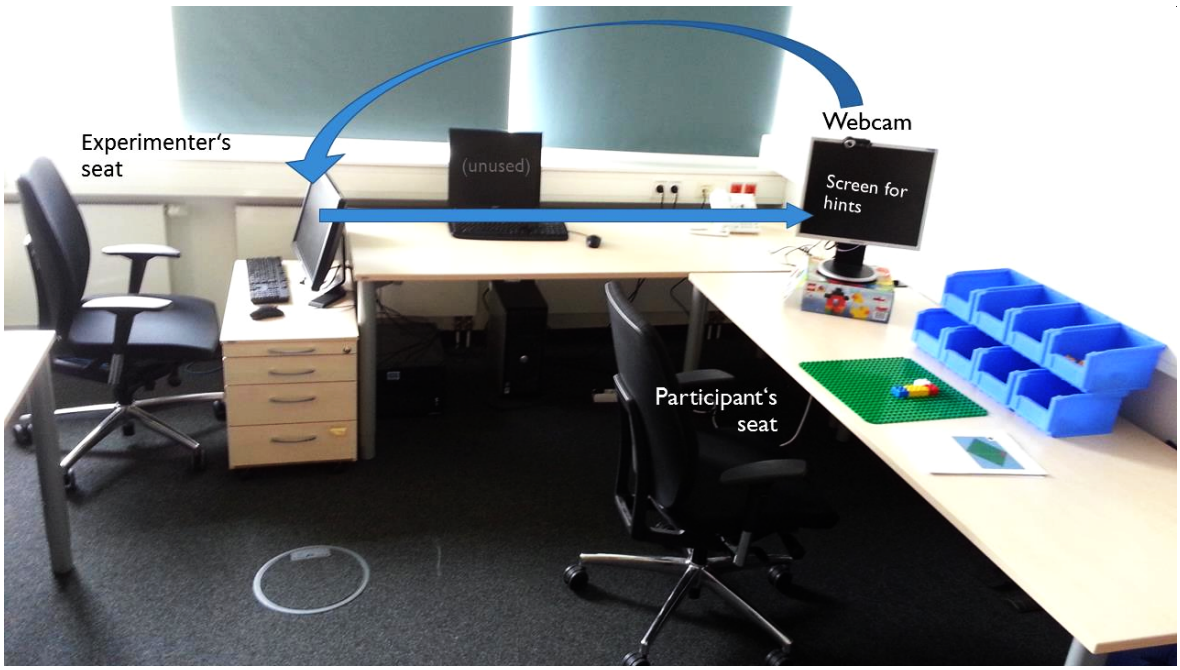

Figure 7.3: **Lab setup** for the Duplo assembly study. A webcam live stream of the assembly area enabled the experimenter to observe participants' actions and intervene on errors by triggering an auditory signal and displaying the correct action on a screen next to the assembly area.

Hettich study

$N_{\mathrm{H}} = 28$ individuals between 23 and 59 years with a mean age of 40 years ($SD = 10.2$) participated in the study. All participants were employees of company Hettich who had been recruited by our contacts and asked to participate on their own volition during their working hours. The majority (75%) of participants were male. Our contacts used their personal knowledge and informed judgment to assign 50% of all participants to the "laypersons" group and the other half to the "experts" group for the assembly task.

7.2.2 Duplo study procedure

First, participants were welcomed and asked to fill out a questionnaire with demographic data and an informed consent for participation, including audio/video recording of their trial. The physical lab setup is shown in Figure 7.3. Participants were seated in front of a table with a large green Lego Duplo base plate and eight blue boxes that each containing a specific type of

Lego Duplo brick with a unique combination of size and color. A computer screen with webcam was placed to their left. The blue boxes were covered by a blanket throughout the experiment except when participants actually needed to use them. During assembly, the webcam recorded the task execution and streamed a live image to the experimenter's screen in the back. This arrangement ensured that participants could not see the experimenter who silently observed their actions during the trial in order to mitigate potential experimenter effects. When assembly errors occurred, the experimenter used the computer to intervene by sending hints to the participants' screen. The procedure of each experimental trial was divided into four steps: SDA-M introduction and pretest, task-related learning phase, SDA-M posttest, and self-reliant task execution. These are subsequently described in more detail.

1) SDA-M introduction and pretest

Participants were instructed how to make decisions during the SDA-M split procedure. Depending on participants' native language, the split instruction either read

> *"Are the depicted steps sequentially associated during assembly, i.e. performed immediately before/after another during task execution?"* (English)
>
> or
>
> *"Sind die dargestellten Aktionsschritte sequentiell zusammengehörig, d.h. werden sie unmittelbar vor- bzw. nacheinander durchgeführt?"* (German)

Printed examples of some assembly actions (with respect to a hypothetical Duplo construction not used in the actual study afterwards) and related SDA-M split decisions (marked as correct or incorrect) were handed out to participants. When they had worked through the examples, three test cases were shown and participants were asked for their decision to verify that they had understood the instructions.

An SDA-M pretest using the QSplit SDA-M software on a tablet computer was then conducted in order to verify that algorithmic assessments of

participants' task-related memory structures reflected their lack of applicable previous knowledge regarding the task structure before they learned about it. A picture of the final result of task execution as shown in Figure 7.1, i.e. the complete target construction consisting of the first 12 bricks from a standardized 16-brick construction by Funk et al. (2015), was briefly shown to participants (for 1 second) prior to the SDA-M split procedure, so they could have recognized it if they had known it and were informed which activity the split procedure refers to. As expected, all participants later confirmed verbally that they did not recognize or know how to build the construction at this point. Each action representation in the SDA-M split procedure described a single assembly step, i.e. placing one brick. The images only displayed the new brick that was to be added in the respective step but not any other bricks that would already have been placed in previous steps (see Figure 7.4 for an example). This simplified type of pictorial action representation was chosen because in most cases showing all previously placed bricks as well would have made it rather trivial to infer the sequential order of placement actions (and corresponding decisions in the split procedure) simply by checking whether the images differed by exactly one brick. In order to ensure consistency and comparability between the experimental phases and groups, action representations for the 12 correct assembly steps as well as for the three wrong actions from the EI group instructions have been incorporated in the SDA-M split procedure, resulting in a total of 15 action representations. This selection was in line with the prevailing approaches in the previous research, which either confined SDA-M split procedures exclusively to representations of actions that constitute correct action sequences for a given task (e.g. Braun et al., 2007; Frank et al., 2016; Jacksteit et al., 2017; Schack & Ritter, 2009; Weigelt et al., 2011) or additionally included representations of a few typical errors (e.g. Hülsmann et al., 2019). Since in principle any kind of brick could have been placed anywhere on the base plate in any step, this confinement of the split procedure strongly violated AMPA's and CASPA's theoretical assumption of completeness.

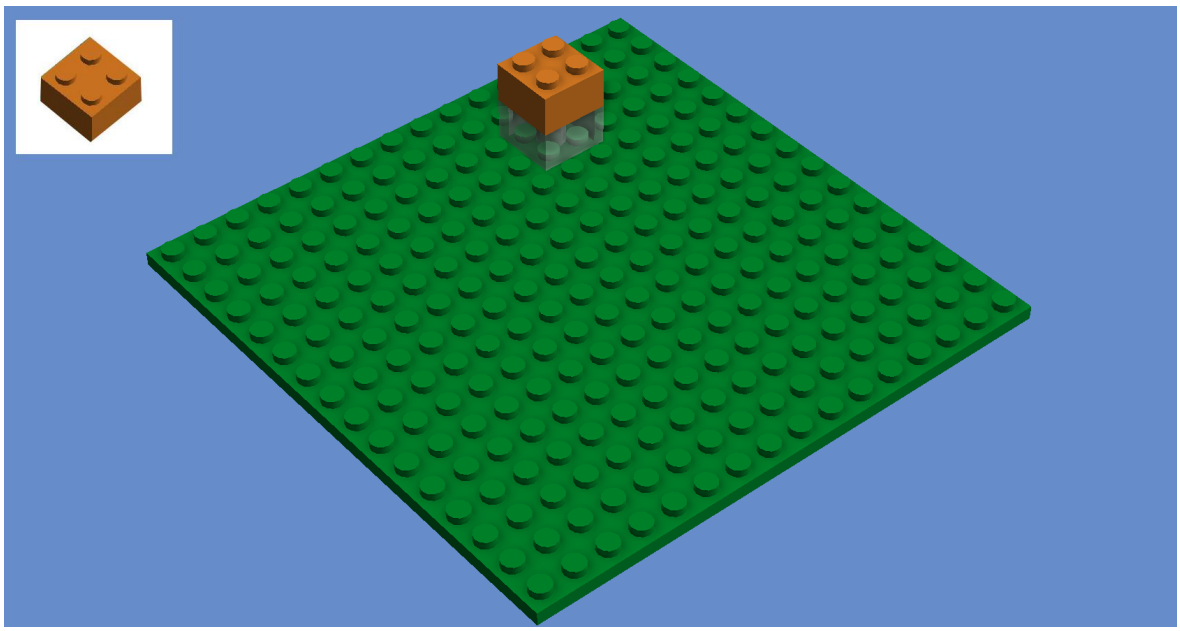

Figure 7.4: **Pictorial representation of a placement action** in the Duplo assembly study. The transparent placeholder brick indicated that the new orange brick must be added on top of another brick at the same X,Y position. The QSplit user interface for the SDA-M split procedure also showed a simple textual description of the action (*"Placing a small orange brick"*).

2) Learning phase

The task then had to be learned by participants so that they would be able to execute it reasonably well. Since the task was previously unknown to participants, this learning phase was obviously necessary to activate or establish some basic task understanding, the related problem solving operators and meaningful task-related mental representation structures in the first place. The contents of the instructions represent an independent variable with two different levels: The CI group (task execution guided by fully correct pictorial step-by-step instructions) and the EI group (task execution guided by partially incorrect pictorial step-by-step instructions with "wrongly" colored bricks in assembly step 3, 8, and 12). This learning material resembled the type of printed step-by-step assembly instructions used by Funk et al. (2015) and was similar to the stimuli used to represent action steps in the split procedure (as in Figure 7.4) but additionally contained all bricks from previous steps (i.e. the entire state of the construction at a specific instant). Participants

were informed that only the relative positions of bricks were actually relevant, not the absolute position related to the green base plate. The learning phase was limited to 4 minutes. Participants could assemble and disassemble the construction and look at the step-by-step instructions as many times as they wanted within this time frame. After the learning phase, participants were instructed to turn away from the assembly area and let the experimenter disassemble whatever they built, replace all bricks to their respective boxes, and cover them with a blanket.

3) SDA-M posttest

Next, the SDA-M split procedure was performed again to update the data about participants' task-related mental representation structures. A picture of the complete designated 12-brick reference construction was again briefly shown to participants (for 1 second) prior to the SDA-M split procedure.

4) Assembly task execution

Participants were then asked to execute the task (i.e. build the designated construction) without guidance, i.e. solely based on their own task knowledge. They were instructed to only touch those pieces they needed to assemble in the current step. All required bricks for the assembly task were arranged in the boxes on the table in front of them. The experimenter supervised the assembly by observing a live camera image. Whenever a participant put his or her hand in a box containing pieces that were not needed in the current step, as well as when they placed a correct brick at a wrong position, this counted as an error. Apart from that, errors were also counted when a participant claimed to not know how to proceed. Whenever a participant made such an error during action execution, the experimenter triggered an assisting hint for the participant which was announced by an audio signal and displayed the correct assembly action for 5 seconds on the participant's screen to their left. This enabled participants to always continue with a correct subassembly at any point within the process.

7.2.3 Hettich study procedure

This study took place in a spacious industrial working environment of company Hettich. Two trials were executed in parallel in different partitions of the hall, each by a dedicated experimenter. The space between the two assembly areas was large enough to prevent participants from directly and deliberately interacting, so they could not assist or copy from each other. However, indirect interference factors such as mutual distraction due to noises during assembly were deliberately left uncontrolled.

First, participants were welcomed and asked to fill out a questionnaire with demographic data and an informed consent for participation. The subsequent procedure of each experimental trial was divided into three steps: Assembly-related instruction, SDA-M introduction and split procedure, and self-reliant task execution. These are subsequently described in more detail.

Figure 7.5: **Participant assembling the Hettich drawer system mockup.** (Photo: Hettich. Used with permission.)

1) Assembly-related instruction

Participants received printed instructions with pictorial and textual descriptions how to assemble a specific drawer system mockup in eleven steps based on educational material from company Hettich. In order to account for participants' considerably differing task-related capabilities and previous knowledge no strict time limit was imposed on the learning phase. Participants were asked to take reasonable time looking through and trying to remember the instructions, and inform the responsible experimenter when they felt ready.

2) SDA-M introduction and split procedure

The SDA-M split procedure was explained by showing participants a special tutorial video included in the QSplit software, which specifies the instructions as follows (in German):

> *Die Software blendet Darstellungen von je zwei Teilschritten der Handlung ein. Sie sollen entscheiden, ob diese Teilschritte bei der Durchführung "direkt sequentiell zusammenhängen" oder nicht, d.h. ob diese unmittelbar vor- oder nacheinander durchgeführt werden. Hierbei spielt keine Rolle, welcher Teilschritt links bzw. rechts angezeigt wird.*

In English this translates to:

> *The software shows representations of two action steps. You shall judge whether these action steps are sequentially "directly associated" during task execution or not, i.e. whether they are executed immediately before or after another. It does not matter which action step is shown on the left side and which one on the right side of the screen.*

The tutorial video continues to illustrate the implications of these instructions using a simple exemplary action sequence for toasting white bread slices and the respective decisions in a corresponding split procedure. Participants were asked to confirm whether had understood these general instructions. After this, they were subjected to an SDA-M split procedure related to the drawer

mockup assembly process. The split procedure incorporated pictorial and textual representations of all eleven assembly steps from the intended action sequence. As an example, Figure 7.2 shows the pictorial representation of the final assembly action *"Fixate Hettich logo at the frame"*.

3) Assembly task execution

Lastly, participants were asked to assemble the Hettich drawer system mockup (Figure 7.5). All required parts and tools were previously placed on a work bench. An experimenter stood by and observed the assembly process. When participants attempted to execute any unintended actions the experimenter took note, intervened verbally by telling them to first reverse the wrong action (if applicable) and helped them execute the correct action instead. This enabled participants to always continue with a correct sub-assembly at any point within the process.

7.3 Results

All four complementary hypotheses could be supported by empiric evidence. As a primary meta result of both studies, the SDA-M-based CASPA algorithm correctly predicted 68.5% to 72.5% of all errors and failures in manual assembly actions, depending on its threshold setting (see Figure 7.6). The subsequent sections first describe additional consolidated results from both assembly studies combined (weighted by the respective numbers of trials), and then the discrete results and ancillary findings for each study individually.

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Balanced accuracy**
AMPA	0.68^{***} a)	0.63	0.69	0.37	0.87	0.66
CASPA_d	0.66^{***} b)	0.68	0.65	0.36	0.88	0.67
CASPA_i	0.65^{***} c)	0.72	0.63	0.35	0.89	0.68

a) $p < 10^{-20}$ b) $p < 10^{-15}$ c) $p < 10^{-14}$

Table 7.2: **Overall results of SDA-M-based error prediction in assembly.**

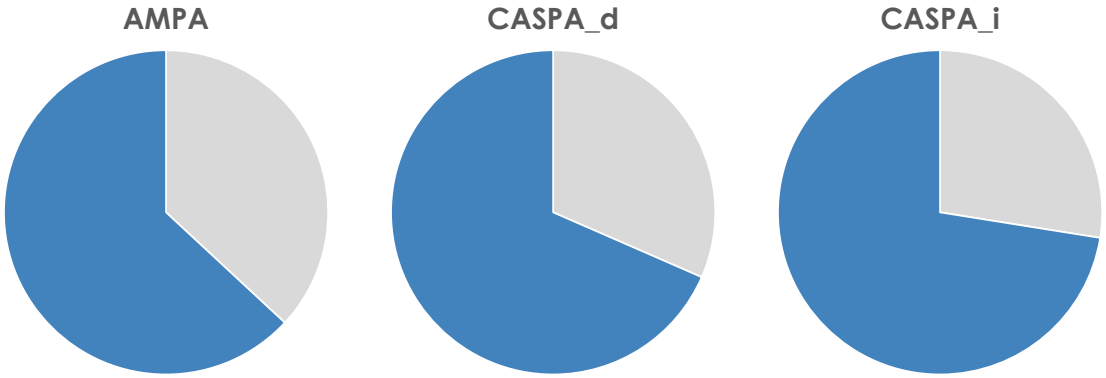

Figure 7.6: **Sensitivity of SDA-M-based error prediction in assembly.** Blue areas indicate the percentage of actual errors that could be correctly detected with AMPA (63%), CASPA$_d$ (68%), and CASPA$_i$ (72%) based on individual SDA-M data in two assembly scenarios.

7.3.1 Consolidated overall results

The overall accuracy, balanced accuracy (Brodersen et al., 2010), and specificity values were comparable for all algorithmic variants (62.6% to 69.4%). One-tailed binomial tests with $H_0 : P(\text{correctPrediction}) \leq \frac{1}{2}$ corroborated that the accuracies of all algorithms were highly significant above chance level (all $p < 0.0001$), providing solid support for the main hypothesis H_1.

Positive predictive values between 35.4% and 36.9% resulting from a relatively low prevalence of errors (149 errors in a total of 676 actions $\Rightarrow P(\text{error}) \approx 22\%$) indicated a notable chance of false alarms, but when the algorithmic SDA-M assessments predicted that an action would be correctly performed without assistance, this was correct in most cases (86.9% to 88.9%). Differences between the three algorithmic variants were marginal. Descriptively, AMPA had slightly higher specificity, whereas both versions of CASPA scored better regarding their sensitivity and negative predictive values.

7.3.2 Detailed study-specific results

The results from both individual studies were similar to the consolidated overall values. In both studies and with all three variants of algorithmic assessments the match between SDA-M-based predictions and actual observations was significantly better than would be expected by chance (see accuracy and

Algorithm	Accuracy	Sensitivity	Specificity	PPV	NPV	Balanced accuracy
AMPA	0.71^{***} a)	0.64	0.75	0.51	0.83	0.69
CASPA_d	0.70^{***} b)	0.69	0.70	0.49	0.84	0.70
CASPA_i	0.70^{***} c)	0.74	0.68	0.50	0.86	0.71

a) $p < 10^{-17}$ b) $p < 10^{-15}$ c) $p < 10^{-15}$

Table 7.3: **Results of error prediction for Lego Duplo assembly.**

Algorithm	Accuracy	Sensitivity	Specificity	PPV	NPV	Balanced accuracy
AMPA	0.63^{***} a)	0.59	0.64	0.17	0.92	0.62
CASPA_d	0.60^{***} b)	0.66	0.59	0.17	0.93	0.62
CASPA_i	0.57^{**} c)	0.66	0.56	0.16	0.93	0.61

a) $p < 10^{-5}$ b) $p = 0.0007$ c) $p = 0.0098$

Table 7.4: **Results of error prediction for Hettich drawer assembly.**

p-values in Table 7.3 and Table 7.4). Descriptively, for each of the three algorithms almost all metrics (except for negative predictive values) turned out slightly better in the Duplo assembly study than in the Hettich drawer scenario.

Auxiliary findings from Duplo study

The complementary analysis of participants' initial task-related memory structures turned out as expected. CASPA estimated an average probability of only 18.7% that participants would have chosen correct actions for building the designated construction before they went through the learning phase for the Duplo assembly task, in contrast to a significantly higher assessed average probability of 57.8% after the learning phase (two-sided Wilcoxon signed-rank test, $W = 2$, $p < 0.0001$). This corroborates hypothesis H_2. While the assessments based on SDA-M measurements *after* the learning phase matched participants' subsequent actual performance significantly *better* than would be expected by chance (see Table 7.3), the assessment of participants *initial* (pre-learning) memory structures matched their actual (post-learning) task performance significantly *worse* than flipping a coin (CASPA_d,

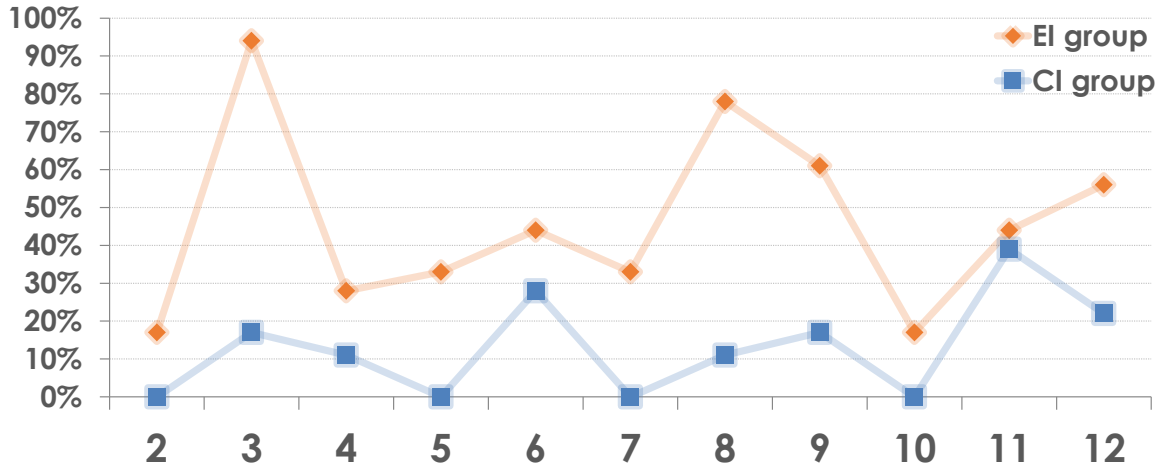

Figure 7.7: **Frequencies of errors in each step of the Lego Duplo assembly task by participant groups.**

two-sided binomial test, $p = 0.00016$). This corroborates the supposition that the automatized SDA-M-based assessments were actually sensitive to changes in participants' task-related memory structures that were presumably caused by the learning phase, which is in line with hypothesis H_3.

As intended by the study design, participants in the Duplo study who received partially erroneous assembly instructions ("EI group") made notably more errors than the correctly instructed participants ("CI group"). Unsurprisingly, the majority of participants in the EI group made mistakes in the three wrongly instructed assembly steps (see steps 3, 8, and 12 in Figure 7.7). They also generally made significantly more errors than the CI group (91 errors vs. 26 errors in a total of 198 attempted action executions; two-sided Mann–Whitney test, $U = 297.5$, $p < 0.0001$), supposedly mainly due to increased levels of cognitive stress and confusion caused by the necessary corrective interventions after they made mistakes. These ancillary findings confirm that the study design worked as intended and successfully induced heterogeneity among participants concerning task-related knowledge and performance in order to yield more meaningful and robust main results.

Auxiliary findings from Hettich study

As a prerequisite for testing hypothesis H_4, the overall performances of laypersons and experts were compared to verify that participants were cor-

	AMPA		CASPA$_d$		CASPA$_i$	
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**
Experts	65.0%	60%	61.4%	60%	57.9%	55%
Laypersons	61.4%	60%	57.9%	60%	56.4%	60%

Table 7.5: **Central tendencies of individual assessment accuracies for Hettich drawer assembly by participant groups.**

rectly assigned to the two groups. It could be confirmed that the experts were actually better at the tested assembly task: In total, the laypersons group made almost twice as many errors as the experts group (21 errors vs. 11 errors in a total of 140 attempted actions for each group). In this context it should be noted that some members of the experts group attempted to partially deviate from the officially defined reference procedure and chose alternative approaches for assembling the drawer. While these attempts may not actually have been erroneous in a practical sense, the study design required them to be treated as such in order to enable proper comparisons. If these alternative approaches had been permitted then the difference between laypersons' and experts' numbers of errors would supposedly have been even more pronounced in favor of the experts.

The accuracies of SDA-M-based assessments were calculated with all three algorithmic variants for each individual participant in both groups. Descriptive statistics (see Table 7.5) and the results of two-sided Mann-Whitney U tests did not indicate that the accuracies of SDA-M-based assessments differed between experts and laypersons for any of the three algorithms (AMPA: $U = 94, p = .87$; CASPA$_d$: $U = 100.5, p = .92$; CASPA$_i$: $U = 93.5, p = .85$). This corroborates hypothesis H_4.

7.4 Discussion

Substantial connections were found between task-related mental representation structures and actual performances in manual assembly tasks. In both conducted studies the majority of human errors, as well as correct action selections, could be properly predicted with computational analyses based on participants' individual SDA-M data. The overall low prevalence of errors led

to comparatively low positive predictive values though. Cognitive assistance systems based on this information should therefore enable proficient users to effortlessly discard unneeded suggestions in case of falsely predicted errors.

The empiric results did not indicate a clear superiority of one algorithmic approach over another. All tested variants (AMPA, CASPA$_d$, and CASPA$_i$), which were based on different cognitive models and parameters concerning action selection mechanisms, performed comparably well. For most practical applications in industrial assembly scenarios, sensitivity would likely be considered the most important metric in order to anticipate as many actual errors as possible. In this regard, the CASPA approach tended to work slightly better than AMPA. The CASPA$_i$ algorithm would have correctly predicted approximately eight out of eleven actual errors in the studies, but the remaining three errors would have been unanticipated. For this reason, practical applications cannot rely solely on this information as a means for preventing all possible errors. This was not surprising from a theoretical point of view for two reasons: First, the assumption of completeness was violated by restricting the number of possible actions that were considered in the SDA-M split procedures. Since split procedures have a time complexity of $\Theta(n^2)$, i.e. the time for completing them grows quadratically as a function of the number of actions, limiting the cardinality of action sets is essential for all practical applications. Second, the task-related mental representation structures retrieved by SDA-M can generally only indicate individual mistakes but not situation- or context-related slips that may arise e.g. due to temporary distractions. For this reason it is also not surprising that the tightly controlled lab study (Duplo assembly) descriptively indicated slightly stronger relationships between assessments of mental representation structures and actual performances than the data acquired in the more unstable surroundings that served as a realistic test environment at company Hettich. However, since the two studies differed not only in terms of environmental controlledness but also several other aspects, a direct comparison for drawing conclusions in this regard is not feasible.

Overall, the results make automatized SDA-M-based assessments of workers' task-related memory structures in manual assembly appear promising as a means for providing individualized on-the-job training or tailoring

cognitive assistance systems to users' personal requirements. The consolidated meta results were based on data from two very different assembly tasks involving a diverse sample of participants in terms of age, sex, educational background, and task-related experience, in order to enhance the robustness and generalizability of the results. However, even if the outcomes of the Duplo assembly and Hettich drawer assembly studies were roughly comparable, it cannot be ruled out that substantially different results would be found in other assembly scenarios and with workers that have other cognitive characteristics than the samples in these studies. Notable limitations of the studies include that no effort was made to manipulate and analyze the influence of factors such as the duration of learning, the time elapsed between learning and task execution, and the number of required assembly actions. Another worthwhile goal for further research would be to investigate the stability or volatility of task-related mental representation structures and corresponding assessments over time depending on the frequency of task executions. Ideally, this could yield practically useful insights about how frequently the SDA-M split procedure needs to be repeated in order to confine violations of the currentness assumption and adequately track workers' learning curves.

技術より心術

Mentality over technique

– Gichin Funakoshi (1938)

Chapter 8.

Prediction of expertise and human error in sequential movements

This chapter is based on:

Strenge, B., Koester, D., & Schack, T. (2020).
Cognitive Interaction Technology in Sport – Improving Performance by Individualized Diagnostics and Error Prediction. *Frontiers in Psychology, 11*, 3641. doi:10.3389/fpsyg.2020.597913

Abstract

The interdisciplinary research area Cognitive Interaction Technology (CIT) aims to understand and support interactions between human users and other elements of socio-technical systems. Impressive developments of new technologies like cognitive robotics, virtual or augmented reality systems, cognitive glasses and neurotechnology settings aroused interest in understanding CIT also in traditionally rather non-technical fields such as sport psychology and movement science. This chapter introduces this ongoing research area and disseminates how automatized analyses of individual mental representation structures based on the algorithms from Chapter 4 could be applied to sequential movements, such as choreographed movement patterns in dance or martial arts, in the context of a new measurement and training

system. Empirical investigations with karate practitioners of different skill levels demonstrate that the SDA-M-based algorithms AMPA and CASPA could generate individualized performance predictions for a movement sequence from the *Kanku-dai kata* (a pre-defined karate movement sequence), which correlated significantly not only with formal expertise (*kyu*/*dan* rank) but also with the actual likelihood of mistakes in action execution. This information could prospectively be used to define individual training goals for deliberate practice and incorporated into cognitive assistance and interaction technology to provide appropriate feedback. Potential benefits of such an assistance system for intermediate and advanced practitioners include more effective and flexible practice, as well as supportive effects, and more flexible training schedules. In order to bring these potential benefits to fruition, the development of technical systems should adhere to process models like *Agile Worth-Oriented Systems Engineering* (AWOSE; see Chapter 9) that explicitly take ethical issues into consideration.

8.1 Introduction

For over a decade numerous researchers from psychology, computer science, engineering, biology, linguistics, and sports science shaped the interdisciplinary field of Cognitive Interaction Technology (CIT) in order to *"generate the scientific insights and the technological basis for creating systems that can interact at various levels of cognitive complexity"* (Ritter & Sagerer, 2009, p. 113). Pursuing the vision of intuitive, human-friendly technology that adapts to users' needs (S. Wachsmuth, Schulz, Lier, Siepmann, & Lütkebohle, 2012) by offering intuitive and personalized support in daily routines (Wrede et al., 2017), CIT comprises research topics such as motion intelligence, attentive systems, situated communication, memory and learning (Schack & Ritter, 2013; I. Wachsmuth, 2008). A major goal is *"to develop memory systems that can approximate some of the key features of human memory, such as flexible association, scalability and learning at different levels"* (Ritter, 2010, p. 230). While classic artificial intelligence concentrates on modeling the mind, CIT research focuses more on interactions that take place in the physical world (S. Wachsmuth et al., 2012) and combines

algorithmic approaches with insights from analyses of human and animal motion to establish "a coherent picture about the internal representation of our movement abilities" (Ritter, 2010, p. 230). On the technical side, CIT combines visualization, sonification, haptic, and augmented reality devices, motion capture, simulated agents in virtual worlds, and attentive user interfaces in novel ways (Ritter, 2010). This led to a broad range of technological advancements such as embodied anthropomorphic robots that can aid humans (Ritter, 2010; S. Wachsmuth et al., 2012), intelligent glasses for cognitive assistance (Essig et al., 2016), and smart environments systems with mobile service robots for ambient assisted living (Wrede et al., 2017).

In contrast, sport psychology was traditionally more concerned with topics like analyzing and improving human performance but started to develop new technologies to support sport performance several years ago (see e.g. Hagan Jr., Schack, & Koester, 2018; Schack, Bertollo, Koester, Maycock, & Essig, 2014; Schack, Hagan Jr., & Essig, 2020; Schack & Ritter, 2013). A main question is how to inform assistance systems about the cognitive background (memory) and motion intelligence (motor skills) of the user. From a traditional cognitive psychology perspective (see e.g. Anderson, 2020), the development of human expertise is commonly characterized by *proceduralization*: The learner integrates declarative knowledge into procedural rule sets so that less declarative memory needs to be used, which reduces brain activation in areas like the hippocampus, prefrontal cortex, and anterior cingulate, and decreases latency. Fitts and Posner (1967) famously described this process as a three-stage model, which transitions from an initial "cognitive stage" to an intermediate "associative stage" and terminates in the "autonomous stage". Research has also found that, while potential performance improvements are limited by factors like musculature and age, the time required for cognitive processing may converge against zero as a power function of practice (Anderson, 2020). This characterization of human expertise development has been challenged by the sport psychological theory of *deliberate practice*, which means engaging in training that is *"focused on improving particular tasks"* and *"involves the provision of immediate feedback, time for problem-solving and evaluation, and opportunities for repeated performance to refine behavior"* (Ericsson, 2008, p. 988). This obviously requires that practitioners are

given specific tasks with well-defined goals (Ericsson, 2007). Purportedly, deliberate practice continually improves performance, because *"expert performers counteract automaticity by developing increasingly complex mental representations to attain higher levels of control of their performance and will therefore remain within the cognitive and associative phases"* (Ericsson, 2008, p. 991).

Based on the CAA-A (see Section 2.1.1), sport psychology researchers described the building blocks and levels of the action system that enable us to control movements such as striking the tennis ball at the right time, or coordinating steps and arm movements in dancing or golf, and demonstrated how the measurement of mental representation can be used for applied work in sport, new pathways in mental training (imagery), and to inform technical systems (Frank et al., 2014; Schack, 2020; Tenenbaum et al., 2009). A highly promising application of interactive technology in sport psychology is to provide helpful assistance to athletes in the context of learning. In coaching, trainees' capabilities to respond to an expert's assistance and the coaching system's ability to activate users' learning potential can be observed (Schack, 2020). Coaching a trainee at different interaction levels while practicing and learning a motor task constitutes an interesting scenario not only for supporting motor learning processes but also to understand the effectiveness of current coaching principles (see also Schack, 2020). Based on mental representation analyses in sport (Schack, 2020; Schack & Hackfort, 2007; Schack & Mechsner, 2006), we investigate how coaching could become more individualized and adaptive in the real world and in Virtual or Augmented Reality settings (Schack et al., 2020). To this extent, it is clearly advantageous for a real or virtual coach to know how mental structures form, stabilize, and change in sports (Schack, 2020). Coaches who possesses such knowledge are better able to address the individual athlete on his or her current level of learning and shape instructions to improve training and performance (Schack, 2020).

In this line of research, numerous studies found that the differing mental representation structures of experts and novices can be measured with SDA-M and influenced by appropriate training (e.g. Frank et al., 2014, 2013; Heinen et al., 2002; Schack, 2004; Schack, Essig, et al., 2014; Schack &

Hackfort, 2007; Schack & Mechsner, 2006). A methodological review and evaluation of research in expert performance in sport by Hodges, Huys, and Starkes (2007, p. 164) noted that the SDA-M method *"is expected to aid in our understanding of the usually nondeclarative motor representations underlying expert performance in fast, complex coordinative actions and in identifying the problems novices encounter in understanding motor problems."* The previous Chapter 4 described advanced algorithms for automatized analyses of task-related mental representation structures based on SDA-M related to action sequences. These algorithmic approaches might be useful as a component of future CIT systems, like cognitive glasses, to measure and improve human performance in sport. In this context, SDA-M and its recent algorithmic extensions could serve as a measurement and assessment tool, and smart glasses or other portable devices could provide corresponding feedback for deliberate practice.

This chapter reports on a first empiric study in karate as a proof of concept for this assessment approach. Subsequently, potential ethical benefits and risks, as well as links to ethical aspects of technical system development, are discussed.

8.2 Methods

Karate practitioners of different skill levels were analyzed regarding a choreographed sequence of distinct movements (karate techniques) from the beginning of the so-called *Kanku-dai kata*. Instructors of the popular *Shotokan* style of karate commonly introduce the *Kanku-dai* at some point during students' preparation for the first *dan* black belt or "master" level. The *Kanku-dai kata* can be understood as a long compilation and rearrangement of subsequences from preliminary *katas*, especially the so-called *Bassai-dai* and *Heian katas*, which should be well-known by then. Therefore, most intermediate practitioners supposedly possess extensive experience with some or all of the preliminary *katas* but have limited, if any, knowledge of the *Kanku-dai*. Even advanced practitioners might commonly fall prey to memory interference effects due to wrong matching and association of the corresponding movement patterns. This constitutes an interesting and challenging scope of

application for analyzing mental representation structures, error prediction and performance assessment. The study focused on the first 17 moves from the beginning of *Kanku-dai* up to the first *Manji-uke* blocking technique.

Statement of ethical approval

The study has been approved by the ethics committee of Bielefeld University in written form according to the guidelines of the German Psychological Society (DGPs) and the Association of German Professional Psychologists (BDP). All participants gave informed and written consent to participate in the study.

8.2.1 Participants

Twelve individuals between 18 and 63 years with a mean age of 30.7 years ($SD = 13.3$) participated in the study. The majority (75%) of participants were male. Some basic experience in karate, as indicated by holding at least the sixth *kyu* rank ("green belt"), was required to enable proper determination of individual techniques. This was necessary since the SDA-M-based analyses in this study were concerned with action selection mechanisms for choosing between different karate techniques within the *kata* sequence. The CAA-A model allocates these mechanisms primarily to the level of "mental control" and the associated "basic action concepts" (BACs) as mental representation units (Schack, 2004, see Table 2.1). The corresponding SDA-M-based analyses in this study were inherently and deliberately indifferent to the quality of individual karate techniques. Therefore, participants had to know and apply these BACs, i.e. execute karate techniques, sufficiently well to allow the experimenter to properly and unambiguously recognize and distinguish them. Table 8.1 shows the exact distribution of participant numbers across formal ranks of expertise. They were reimbursed for their time with 5 Euros in cash.

8.2.2 Procedure

First, participants were welcomed, asked to give informed consent to participation, and provide demographic data, as well as their degree of formal

Rank	6th *kyu*	5th	4th	3rd	2nd	1st *kyu*	1st *dan*	2nd *dan*
No. of participants	2	3	1	1	1	0	1	3

Table 8.1: **Formal expertise of participants in karate.** Note that expertise increases from left to right, because *kyu* ranks traditionally decrement from eighth (beginner) to first *kyu* (advanced student), whereas the subsequent *dan* ranks (”master level”) are counted upwards from 1st *dan*.

expertise in karate. The following proceedings of each trial could be divided into three consecutive phases:

1) Recapitulation and learning phase

A brief recapitulation of preliminary *katas* served both as a physical warm-up and cognitive trigger for activating relevant memory structures. This included the *Heian Nidan*, *Heian Yondan*, and *Bassai-dai*, which contain similar or identical parts as *Kanku-dai*, each from beginning until the first occurrence of a *kiai*[1]. Participants who had already been tested in a given *kata* as part of an official examination for their *kyu* or *dan* grade were merely asked to demonstrate it once, in a calm and serene manner, without further guidance. The remaining preliminary *katas* were at least once roughly synchronously executed by the participant and the experimenter as an instructor. If participants made mistakes or struggled noticeably the execution was repeated up to two times. Afterwards, a video was shown of the *Kanku-dai* sequence performed by Master Hirokazu Kanazawa (10th *dan* black belt; †8 December 2019). Participants were then rudimentarily taught to execute this sequence by following the moves in rough synchrony with the experimenter. The number of repetitions depended on formal expertise ranks: Relative beginners (eighth to fifth *kyu*) executed the sequence twice, advanced students (fourth to first *kyu*) executed it once, and black belts did no physical execution at all. The video of the *Kanku-dai* sequence was then shown a second time to finalize the learning phase.

[1]The *kiai* is a short shout that is uttered when performing distinct moves in karate. The correct execution of *katas* usually requires *kiais* at certain specified points.

2) SDA-M introduction and split procedure

The SDA-M split procedure was explained by showing participants a special tutorial video included in the QSplit software, which specifies the instructions as follows (translated from German to English):

> *The software shows representations of two action steps. You shall judge whether these action steps are sequentially "directly associated" during task execution or not, i.e. whether they are executed immediately before or after one another. It does not matter which action step is shown on the left or on the right side of the screen.*

The tutorial video continues to illustrate the implications of these instructions using, as a simple example from daily life, an action sequence for toasting white bread slices and the respective decisions in a corresponding split procedure. Participants were asked to confirm whether they had understood these general instructions. After this, they were subjected to an SDA-M split procedure, which incorporated still images of the first 17 techniques of the *Kanku-dai kata* and corresponding textual descriptions. As usual in karate, Japanese terms were used to denote the techniques.

3) Movement sequence execution test

Lastly, participants' capability to freely execute the *Kanku-dai* movement sequence was tested. Participants started the *kata* with their back towards the experimenter, so they could not see the experimenter during the movement sequence execution. The experimenter observed the execution and intervened when errors occurred. In this case the experimenter told participants to freeze in their current position, walked in front of them, and demonstrated the correct technique. Participants should then reverse their previous (wrong) action and continue with the correct execution. This intervention procedure was beforehand explained and demonstrated. Importantly, merely slightly inaccurate action executions were ignored as long as the correct technique was still clearly recognizable. Only wrongly chosen techniques were counted as errors and corrected.

8.2.3 Data analysis

All SDA-M procedures were executed with the *QSplit SDA-M Suite* v1.6 for Windows. This included the split procedure and the usual data normalization, scaling, clustering and invariance analysis steps (see Chapter 3), as well as advanced analyses using the AMPA and CASPA algorithms (see Chapter 4). Generally, the available data were analyzed on two different levels:

First, on the level of individual karate techniques, the algorithmic predictions by AMPA, CASPA$_d$ and CASPA$_i$ for each action of every participant were compared with the corresponding outcomes during actual execution. For this purpose, several standard metrics for the evaluation of binary classifiers were used. In this context a "true positive" case was counted when the algorithmic analysis predicted an error and this error actually occurred.

Second, participants overall performances, i.e. total numbers of correct actions, and their formal expertise ranks were compared with different SDA-M-based measures, which aim to reflect the overall suitability of individual mental representation structures for the movement task. The two "traditional" measures, the invariance measure λ and the ARI, are based on SDA-M clustering results (see Section 3.1.5). This implies they require a reference structure for comparison, e.g. from one or multiple domain experts. In the present study, an ideal reference structure for this purpose was established by perfectly associating the action representations that exactly precede or follow each other in the movement sequence.

In addition to these two established measures (λ and ARI), CASPA$_m$ is newly introduced as an advanced alternative. It represents the arithmetic mean over all likelihoods of successful action selection during the whole sequence of movements as predicted for an individual by the CASPA algorithm. Formally, if n is the number of actions in the designated action sequence (here: $n = 17$) and p_i the probability of correct action selection for a given participant after executing a previous action a_i as estimated by CASPA, then CASPA$_m$ is defined as follows:

$$\text{CASPA}_m := \frac{1}{n-1} \sum_{i=1}^{n-1} p_i; \text{CASPA}_m \in [0, 1] \tag{8.1}$$

This value can also be interpreted as an overall estimate of the expected prob-

Algorithm	Accuracy	Sensitivity	Specificity	PPV	NPV	Balanced accuracy
AMPA	0.71^{***}	0.55	0.75	0.29	0.90	0.65
CASPA_d	0.69^{***}	0.74	0.68	0.31	0.93	0.71
CASPA_i	0.69^{***}	0.77	0.67	0.31	0.94	0.72

*** $p \ll 10^{-5}$

Table 8.2: **Detailed results of SDA-M-based error prediction in the *Kanku-dai* sequence.**

ability of correct action selection for a randomly chosen situation within the sequence.[2] CASPA_m has the advantage over previous alternatives (the invariance λ and ARI) that it does not require an explicit reference structure. CASPA_m also inherits a notable limitation of the CASPA algorithm though: It is only applicable to SDA-M data sets related to action sequences that have no temporal overlap between the actions. Therefore, it cannot generally replace λ and ARI for arbitrary SDA-M application scenarios if this condition is not satisfied.

8.3 Results

Substantial, albeit imperfect, matches between algorithmic analyses of participants' mental representation structures and their actual accomplishments while executing the movement sequence were found.

Detailed metrics for the performance of AMPA, CASPA_d (using the default threshold of 0.5), and CASPA_i (using an empirically informed threshold of 0.6207), with respect to predicting participants individual likelihood of making mistakes at the level of each individual action (i.e. discrete karate techniques) are shown in Table 8.2. An overall relatively low prevalence of errors (31 errors in a total of 192 actions $\Rightarrow P(\text{error}) \approx 16\%$) caused a salient discrepancy between positive and negative predictive values (PPV and NPV). However, the prevalence-independent measures of sensitivity and specificity were rather close to each other. From an applied perspective sensitivity mat-

[2]This formulation assumes that there is only one correct action sequence to achieve the goal. The case of multiple different correct sequences would require slightly more complex calculations, involving a weighted arithmetic mean that weights the estimated probabilities of correct action selection for each situation with the joint probabilities of having previously chosen exactly the actions needed to get into that situation.

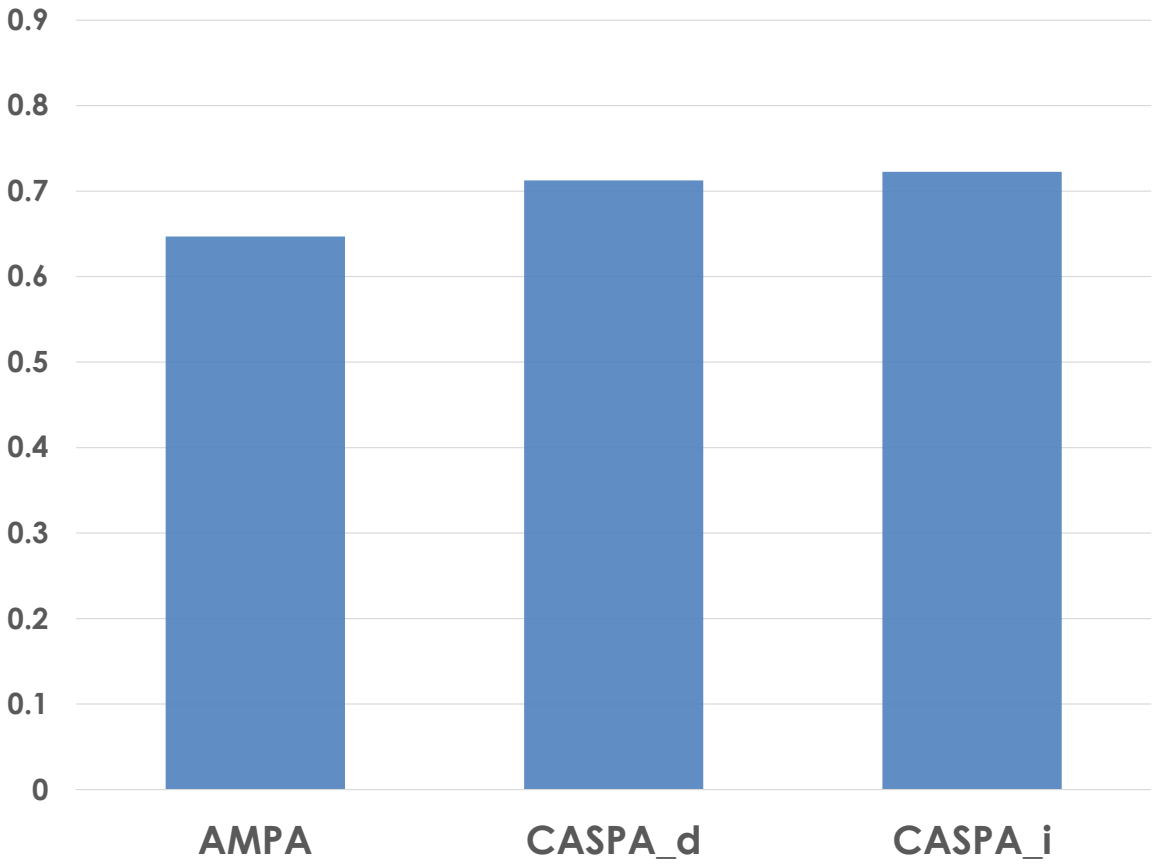

Figure 8.1: **Balanced accuracies of different SDA-M-based algorithms** for error prediction in the *Kanku-dai* sequence.

ters for recognizing as many of the practitioners weak points as possible, whereas specificity helps focusing on these issues instead of unnecessarily practicing parts they already mastered. The CASPA$_i$ algorithm achieved the best results among the different algorithmic variants in terms of balanced accuracy (Figure 8.1), which represents the arithmetic mean of sensitivity and specificity values (Brodersen et al., 2010).

Table 8.3 shows the correlations (using Spearman's rank-order correlation coefficient ρ) between participants' formal expertise (*kyu*/*dan* rank), their actual performance in the *Kanku-dai kata* execution test (i.e. number of correctly chosen techniques), the conventional SDA-M measures for assessing the invariance and similarity of individual mental representation structures to an ideal reference structure (Lander's λ and ARI)[3], and the newly proposed

[3]Lander's λ and ARI were both calculated from SDA-M clusterings with a significance level of $\alpha = 1\ \%$.

	Expertise	**Performance**	**CASPA**$_m$	**Invariance** λ
Performance	.84***			
CASPA$_m$	.80$^{**\ b)}$	.88***		
Invariance λ	.66$^{*\ a)}$	.79$^{**\ c)}$	.80$^{**\ a)}$	
ARI	.44	.65$^{*\ b)}$	.49	.75$^{**\ d)}$

*** $p < 0.001$ ** a) $p = 0.0016$ ** b) $p = 0.0018$ ** c) $p = 0.0022$ ** d) $p = 0.0047$
* a) $p = 0.02$ * b) $p = 0.023$

Table 8.3: **Correlations between formal expertise, actual performance, and SDA-M-based assessment metrics.**

CASPA$_m$ measure. All three SDA-M-based assessment metrics (Lander's λ, ARI, and CASPA$_m$) showed significant and strong positive correlations with participants actual performances. CASPA$_m$ and Lander's λ also correlated significantly and strongly with formal expertise ranks. The differences between Lander's λ, ARI and CASPA$_m$ were statistically insignificant (using Fisher z-transformation for comparison of the correlation coefficients). Descriptively, CASPA$_m$ showed the strongest correlations with performance and expertise among all three SDA-M-based metrics. Furthermore, CASPA$_m$ values showed higher correlations with actual performance than formal expertise ranks did.

8.4 Discussion and ethical considerations

Deliberate practice has generally been accepted as an important factor for developing expertise, especially in sports, even though the specific extent of its impact on performance remains a subject of debate (cf. Anderson, 2020; Ericsson, 2008; Macnamara, Moreau, & Hambrick, 2016). By definition, deliberate practice requires that a coach or trainer sets specific individual training goals and provides feedback to practitioners. This may constitute a blocking obstacle when no coach is available, e.g. during travel or exercise at home. Motivated by prior research results and applications of the SDA-M method the present study investigated whether automatized SDA-M-based assessments could serve as an approximate technical substitute for the role

that human coaches fulfill in deliberate practice. This included identifying potential issues and assessing a practitioner's overall competency with respect to specific movement sequences to derive feasible training goals.

Albeit preliminary due to a limited sample size, the empiric results are highly promising: SDA-M-based algorithms reached accuracy values that were highly significant above chance level and correctly predicted up to 77% of all actual errors in action selection during the tested karate movement sequence. Furthermore, SDA-M-based measures for assessing the overall suitability of participants' individual mental representation structures, especially the newly proposed CASPA$_m$ metric, correlated significantly and strongly with karate practitioners actual performances.

The present study focused on choosing correct movements, not on improving individual actions' execution quality. Arguably, assisting deliberate practice on the level of basic action selection rather than the level of atomic action features seems especially helpful for intermediate and advanced practitioners, since Ericsson (2008, p. 991) noted that after sufficient practice *"the aspiring expert performers become able to monitor their performance so they can start taking over the evaluative activity of the teacher and coach. They acquire and refine mechanisms that permit increased control, which allow them to monitor performance in representative situations to identify errors as well as improvable aspects."* While this kind of self-monitoring might work well for recurring basic actions, like well-known karate techniques, it cannot prevent mistakes in insufficiently practiced action sequences.

A notable limitation of the currently available algorithms for automatized SDA-M-based assessments and error predictions is that they require a predefined, limited set of correct action sequences in terms of basic actions. This makes them potentially applicable not only to martial arts forms and dance choreographies but also to opening sequences in chess or real-time strategy games (B. Strenge *et al.*, unpublished) and other fixed sequences of basic actions that do not overlap in time. However, they cannot readily be applied to more dynamic, impulsive and spontaneous situations in sports and training that do not satisfy these requirements.

Future research could focus not only on replicating the current study's findings with more extensive and heterogeneous participant samples and

other sample applications but also investigate the long-term applicability and usefulness of the automatized assessment approaches. A major research and development objective could be to build an assistance system and empirically test its impact on the quality and efficacy of deliberate practice compared to unassisted training and/or traditional coach interaction. A mobile CIT assistance system, e.g. based on smart glasses, could use the information from SDA-M-based analyses to suggest training goals, provide athlete- and sport-specific feedback, and track practitioners' learning curves in terms of developing task-related memory structures over time. Such a system would enable intermediate practitioners to engage in deliberate practice of action sequences anywhere anytime instead of requiring personal contact with their coaches. Furthermore, intelligent smart glasses could enable remote observation or assistance (e.g. transferring the video to coaches and allowing them to use a salient pointer to help athletes focus on relevant cues), as well as new forms of training, such as displaying distracting stimuli in the glasses in order to simulate different training conditions or environments (see Schack, 2020). Arguably, this would entail a broad range of ethically relevant aspects:

- Greater independence from organizational structures like sports clubs,
- less time spent and environmental damage due to regular traveling,
- more flexible training schedules,
- better opportunities for independent adjustment of repetitions in deliberate practice, and
- prevention of potential embarrassment due to the observation of ones mistakes by other people.

With respect to the current situation concerning the ongoing COVID-19 pandemic and impending climate catastrophe, one might add that special circumstances make many of these aspects all the more relevant and pressing issues.

A CIT system would need to be developed with the aforementioned and other ethical aspects in mind to ensure that the potential benefits actually

come into effect. Despite all the new possibilities opened up by the application of new technologies in sport science there are also many challenges that have to be considered: New technologies allow the recording and storage of detailed user-specific data. Privacy issues and other ethical, legal and social implications (ELSI) are becoming more and more important and are seen as essential considerations with respect to technological developments. Therefore, the technical development process should adhere to specific rules regarding the inclusion of ethical issues. This is especially important in contemporary agile development settings that are characterized by transient requirements definitions and short-term prioritization of features. Specialized system design methodologies like *Agile Worth-Oriented Systems Engineering* (AWOSE; see Chapter 9) define methods and processes for this purpose. Finally, a long-term study could verify which (if any) ethically relevant benefits actually arise from using such a system.

In a related research direction, which could be interesting for anticipatory systems in sports and medicine, researchers tried to support everyday activities with assistive glasses. This endeavor, called ***Project ADAMAAS*** (*"Adaptive and Mobile Action Assistance"*), focused on the development of a mobile adaptive assistance system in the form of intelligent glasses, which provides unobtrusive, anticipatory, and intuitive support in everyday situations (see Figure 8.2 and Essig et al., 2016). The ADAMAAS system aimed to identify problems in ongoing action processes, react to mistakes, and provide context-related assistance via textual, pictorial, or three-dimensional virtual elements superimposed on a transparent display. For this purpose, Project ADAMAAS investigated the integration of mental representation analysis, eye tracking, physiological measurements (e.g. pulse and heart rate variability), computer vision (i.e. machine learning techniques for object and action recognition), and augmented reality with modern diagnostics and corrective intervention techniques. Major perspectives that distinguish the ADAMAAS system concept from stationary diagnostic systems and conventional head-mounted display systems include the ability to react to errors in real-time, provide individualized feedback for action support, and learn from the individual behavior of the user. The following Part II of this thesis presents the human-centered methodology and related proceedings of this project.

Figure 8.2: **Illustration of the ADAMAAS system concept** using AR instructions to assist a baking task. (Photo: CITEC. Used with permission of Thomas Schack.)

Part II

Developing a User-Adaptive Cognitive Assistance System

Preamble

From May 2015 to April 2018, the German Federal Ministry of Education and Research (BMBF) funded Project ADAMAAS at Bielefeld University's Center of Excellence in Cognitive Interaction Technology (CITEC). The goal of this project was to use smart glasses and related augmented reality (AR) setups in combination with eye tracking and psychological measures to provide cognitive assistance for elderly and people with particular handicaps in daily living activities, education or work tasks.

Major parts of Project ADAMAAS have been concerned with the development of 3D AR visualizations and interaction, eye tracking, and machine learning for computer vision and action recognition. I contributed to these parts to a certain extent, most notably the AR user interface concepts, in close collaboration with the responsible colleagues. Another major part –primarily my part– was to investigate the potential and possibilities for integration of automatized SDA-M-based assistance approaches, which have been introduced in Part I of this thesis, into the technical system.

I was also responsible for the project's user-centered engineering methodology and evaluation. Since the project's primary target groups comprised people who are particularly vulnerable, a careful and systematic consideration of ethical issues was an important aspect of the research and development proceedings. To this end, a stakeholder-centered process model for integrating usability engineering methods with ethical analyses and issues into agile development activities has been established in Project ADAMAAS. The following Chapter 9 describes this process model. Subsequently, Chapter 10 presents a selection of related user-centered design activities, as well as complementary approaches to usability- and worth-related system evaluation.

“Do the right thing”

– Alphabet Inc. (2017)

Chapter 9.

The AWOSE development process

This chapter is based on:

Strenge, B., & Schack, T. (2019).
AWOSE - A Process Model for Incorporating Ethical Analyses in Agile Systems Engineering. *Science and Engineering Ethics, 26*(2), 851-870. Springer Nature. doi:10.1007/s11948-019-00133-z

Abstract

Ethical, legal and social implications are widely regarded as important considerations with respect to technological developments. Agile Worth-Oriented Systems Engineering (AWOSE) is an innovative approach to incorporating ethically relevant criteria during agile development processes through a flexibly applicable methodology. First, a predefined model for the ethical evaluation of socio-technical systems is used to assess ethical issues according to different dimensions. The second part of AWOSE ensures that ethical issues are not only identified but also systematically considered during the design of systems based on information and communication technology. For this purpose, the findings from the first step are integrated with approaches from worth-centered development into a process model that, unlike previous approaches to ethical system development, is thoroughly compatible

with agile methodologies like Scrum or Extreme Programming. Artifacts of worth-centered development called Worth Maps have been improved to guide the prioritization of development tasks as well as choices among design alternatives with respect to ethical implications. Furthermore, the improved Worth Maps facilitate the identification of suitable criteria for system evaluations in association to ethical concerns and desired positive outcomes of system usage.

9.1 Motivation

The discovery of nuclear fission by German scientists Otto Hahn and Fritz Strassmann during World War II led to research on nuclear chain reactions culminating in the creation of the first nuclear weapons by the U.S. during the Manhattan Project. A few decades later, after school shootings around the millennium change, authorities were quick to allege that computer games like Counter-Strike had a negative psychological impact on the killers, despite a glaring lack of scientific evidence supporting these claims. More recently, the widespread use of (web-based) social networks not only raised privacy concerns but also created unwanted phenomena like “cyberbullying” or “cyberharassment”. All of these chains of events make it abundantly clear that scientists and engineers are well advised to assess long-term consequences of their research and development projects carefully. In most cases, even the direst worst-case impacts on society may not match the potential of weapons of mass destruction. Nonetheless, the current consensus among researchers is that ethical, legal and social implications (ELSI) are an important aspect of all science and engineering endeavors. However, as discussed in in Section 2.3, few guidance has been offered regarding approaches to systematic handling of ethical issues during actual development processes for systems based on information and communication technology (ICT), especially during the day-to-day work in agile development processes that are characterized by transient requirement definitions and limited overall predictability. This constitutes a pressing issue since an absence of explicit ethical considerations may lead to suboptimal adoption of new technologies such as intelligent assistive systems (Ienca et al., 2017). There are two main issues to solve:

1. How to identify and assess ethical implications, and
2. how to handle these during system development.

This chapter proposes a structured approach to filling the prevailing gaps concerning agile development processes by incorporating ethically relevant criteria through a flexibly applicable methodology called Agile Worth-Oriented Systems Engineering (AWOSE). The AWOSE methodology is based on preliminary work from a computer science master's thesis (Strenge, 2013) and has been extended and refined during Project ADAMAAS (Essig et al., 2016). Throughout this chapter, the core concepts of AWOSE shall be illustrated by examples from its application in Project ADAMAAS.

Overview

In AWOSE's first part, a multi-dimensional model for the ethical evaluation of socio-technical arrangements (MEESTAR; Manzeschke, Weber, Rother, & Fangerau, 2015) is used to identify and assess ethical issues on an individual, organizational and social level, as well as according to a standardized set of dimensions, such as privacy, participation or safety. Each potential issue's severity is evaluated according to a four-level scale that ranges from "completely harmless" to "should be opposed from an ethical viewpoint". As a result, detailed information about relevant ethical issues regarding the socio-technical system is gained.

The second part of AWOSE ensures that these ethical issues are not only identified but also adequately considered during system development process by integrating the MEESTAR-based analyses with approaches from worth-centered development. Special artifacts from the worth-centered development methodology called Worth Maps have been extended and improved to combine project management tools and engineering methods, which guide the regular prioritization of development tasks as well as systematic choices among design alternatives with respect to ethical implications. Furthermore, the improved Worth Maps facilitate the identification of suitable criteria for system evaluations in association with ethical aspects and explicitly relate these to desired user experiences and positive outcomes of system usage.

Finally, this chapter presents a process model for structuring and organizing both parts of the AWOSE methodology, which combines user experience, engineering, and ethical assessments, and is compatible with well-known agile methodologies like Scrum or Extreme Programming.

9.2 Identification and assessment of ethical issues

The first part of AWOSE can start whenever a suitably representative description of the technology, technical system, or product has been developed. (In the following, this chapter will refer to a "system" being developed, and the respective "system vision", but the statements generally hold for technologies and product developments as well.) The description may have any form, e.g. a simple textual description, graphical sketches, diagrams, mockups, or a usable prototype. The degree of how detailed the description should be poses a trade-off, as is the case for many other human-centered system design methods: The less detailed the description, the less reliable will any assessment be that is based on it. The more detailed the description, the fewer degrees of freedom may remain to adjust the design. In general, it is advisable to start with an early version and continually re-iterate the assessment process as deeper knowledge regarding aspects and features of the system and its environment, including users and other stakeholders, becomes available.

In order to identify ethical issues, MEESTAR, a "multi-dimensional model for the ethical evaluation of socio-technical arrangements" created by German philosopher, theologian and anthropology professor Manzeschke et al. (2015), is used. MEESTAR was originally developed for analyses of "age-appropriate assisting systems", which comprise a broad range of socio-technical systems that are supposed to be used primarily by elderly people to help them live autonomously in their own homes. However, it has also been applied in contexts such as assistance for young people with disabilities, telemedicine, and systems for working environments (Manzeschke, 2015). MEESTAR provides a reference framework to structure discussions about system-related ethical aspects, ideally in the form of interdisciplinary workshops, with respect to a set of predefined dimensions (Figure 9.1). As a preparatory step, the model and the system description, including the in-

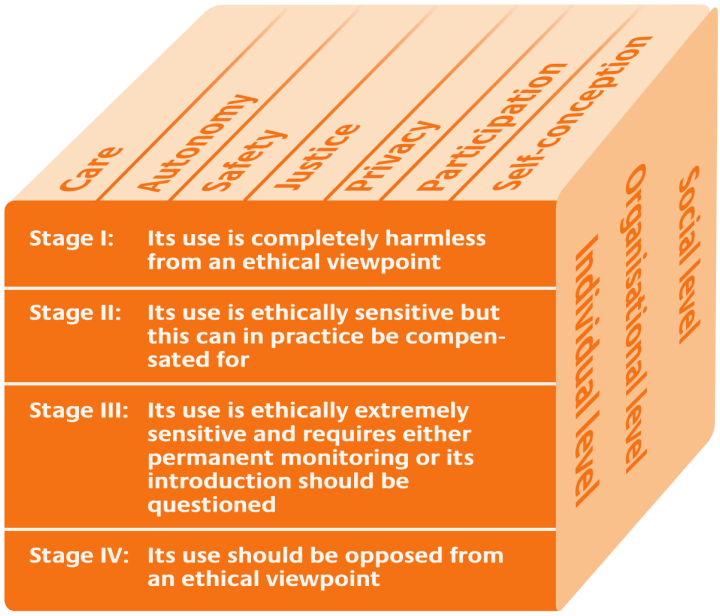

Figure 9.1: **MEESTAR.** The multi-dimensional model for the ethical evaluation of socio-technical arrangements from Manzeschke et al. (2015) is used as the first part of AWOSE.

tended context of use, are presented to all workshop participants. MEESTAR then requires systematic consideration of ethical issues related to seven dimensions (care, autonomy, safety, justice, privacy, participation and self-conception) on an individual, organizational, and social level.

The seven ethical dimensions have been derived from theoretical ethical work as well as a series of qualitative interviews (Manzeschke et al., 2015). These dimensions are not meant to serve as guidelines from which ethical judgments could be derived but to help evaluators “to identify and allocate one or more ethical issues in an actual scenario” (Manzeschke et al., 2015, p. 14). Extensive definitions and corresponding examples for all seven dimensions have been provided by Manzeschke et al. (2015). Nevertheless, there is certainly some degree of overlap and fuzziness concerning the mapping of identified ethical issues to these dimensions. This does not seem to compromise the usefulness or applicability of the model but rather to foster prolific discussions and reflection of each issue. Actually, supporting brainstorming and discussion is the main purpose of referring to MEESTAR’s dimensions, whereas structuring the resulting output may be considered a secondary benefit. Generally, it makes little sense to understand any ethical ”dimensions” or

”values” as objective, absolute and stable constructs for a variety of reasons. Due to the associative nature of human cognition, everyone will understand a term like ”privacy” or “justice” in a slightly different way (see also Umbrello, 2019), and morality changes over time and interacts with technology (Boenink, Swierstra, & Stemerding, 2010). It is therefore highly important that a sufficiently large and (cognitively) diverse set of people participate in the MEESTAR workshops. In a similar vein, MEESTAR's co-creator Weber (2018) acknowledged that the proposed dimensions, combined with preconceptions regarding their meaning, might influence workshop participants and their judgment. He suggested that VSD methods and literature could be used to “systematically identify moral dimensions for MEESTAR” related to a specific project context (Weber, 2018, p. 260). However, he also noted that many projects might not have enough resources to do so. Furthermore, MEESTAR's default dimensions can be mapped to the four basic principles of biomedical ethics (i.e. autonomy, beneficence, non-maleficence, and justice; Beauchamp & Childress, 2006), which have been widely used in ethical evaluation methods (Reijers et al., 2017; Weber, 2018). In order to support its agile orientation, AWOSE uses the abovementioned seven dimensions by default, but the methods for identifying “worth”, which are described in section 9.3.1, as well as other appropriate sources, can be used to inform and adjust this set if necessary. This seems especially important if the developed system's properties deviate significantly from the properties of age-appropriate assistance systems that MEESTAR's default dimensions primarily aim to cover.

9.2.1 Consideration of environmental and nature-related factors

Generally, MEESTAR focusses on addressing ethical concerns related to the wellbeing of human stakeholders. However, issues related to other lifeforms (plants, animals, etc.) are not at all explicitly covered. The analyses should therefore be broadened to include nature-related implications. Nowadays, perceptions and appreciation of non-human life differ widely, ranging from demanding equal treatment of all forms of life, to claims that only human requirements matter. As Steven Umbrello, Managing Director at the Institute

for Ethics and Emerging Technologies, noted: *"Contemporary scholarship on metahumanisms, particularly those on posthumanism, have decentered the human from its traditionally privileged position among other forms of life"* (Umbrello, 2018, p. 3). A consideration of nature-related aspects tends to increase awareness of issues that are undoubtedly as important for the long-term development of human societies as they are for earth's overall ecosystem, e.g. issues related to sustainable production, operation, and maintenance of systems.

9.2.2 Referring to individual, organizational, and society levels

Depending on the total number of workshop participants, the group can be split into subgroups tasked with working on the individual, organizational, or society level. From a user-centered design perspective, the individual level may be considered the most important, but Manzeschke et al. (2015, p. 20) argued that not just individuals have to be responsible for their actions but also corporative entities such as companies, and that a social level of responsibility must be discussed as well. While the relevant organizations and societies are usually identified with relative ease, in the frame of AWOSE a meaningful reference to individuals must be established using specific stakeholder models. Arguably, Personas (Cooper, 2004; Pruitt & Adlin, 2006) are most suitable for this task due to a distinct set of properties:

- they constitute generalizations from real individuals such that a small set of Personas represents large groups of users and other relevant stakeholders,
- they are highly detailed and can be specifically based on relevant types of data from market or user research, and
- they effectively exploit well-developed human capabilities such that designers and developers can easily extrapolate the persona descriptions to infer likely behaviors of the represented "persons" in a given situation (Pruitt & Grudin, 2003).

In order to ensure objectivity, Persona descriptions should be derived from actual data regarding relevant stakeholder properties in a systematic and trace-

able way, e.g. using descriptive statistics and/or approaches based on principal component analysis or factor analysis (McGinn & Kotamraju, 2008; Miaskiewicz, Sumner, & Kozar, 2008; Sinha, 2003), i.e. a reduction of high-dimensional data spaces (e.g. from questionnaires) to a smaller set of uncorrelated linear combinations of the original properties. However, in absence of applicable data, so-called "ad-hoc Personas" (Norman, 2004), i.e. fictive descriptions of hypothetical stakeholders, can still be useful to make explicit statements and reach consensus about the targeted stakeholder groups instead of nontransparent implicit assumptions. In the ADAMAAS project, survey-based data about stakeholder characteristics could be acquired for two of three application scenarios in order to derive Persona descriptions based on statistics. For the remaining scenario, ad-hoc Personas have been created based on researchers' observations and assumptions and then handed over to the application partner's human resources department for validation. In all of these cases, "primary" Personas represented potential user groups, while "secondary" Personas represented indirectly affected stakeholders (e.g. users' supervisors or managers). All Persona descriptions were then printed and handed out to MEESTAR workshop participants.

9.2.3 Assessment of ethical sensitivity

The final step of MEESTAR consists of an evaluation of each identified ethical issue on a scale with four levels, ranging from *"completely harmless"*, *"ethically sensitive"* and *"extremely sensitive"* to *"should be opposed from an ethical viewpoint"* (Manzeschke et al., 2015, p. 14). It is important to note that each of these assessments is explicitly related to 1) one of the seven ethical aspects, 2) a specified individual, organization, or society, and 3) a specific timeframe. In AWOSE, the latter is by default implicitly defined as a snapshot of the current reality at the instant when the assessment takes place, but it may be worthwhile to consider the expected impact of foreseeable developments, especially for upcoming technologies and products with a prolonged lifespan. Since MEESTAR-related analyses in AWOSE are supposed to be conducted by an interdisciplinary group (e.g. researchers, engineers, potential users, practitioners and domain experts with different backgrounds), in many cases

the initial judgments regarding each issue's ethical sensitivity may vary. The proper way of resolving these situations obviously poses an ethical question in itself, which the original publications on MEESTAR did not cover. The pragmatic solution in AWOSE is to try first to reach a consensus on the sensitivity through discussion. If this fails, the highest severity rating chosen by any member of the interdisciplinary group is selected, i.e. the goal is to err on the side of caution.

While the AWOSE methodology requires at least one MEESTAR workshop as soon as the system vision is available, it is often advisable to schedule several iterations and update the list of ethical issues over time as more and more is known about the system and its context of use. During the ADAMAAS project's three-year funding period, five half-day MEESTAR workshops have been organized with an average of nine to ten participants, including representatives of the project partners and stakeholders. Since ADAMAAS could be considered as an age-appropriate assistance system, MEESTAR's default dimensions were used. Retrospectively the initial list of relevant ethical issues had converged towards a reasonably stable set after the third workshop.

Since its creation in 2013, MEESTAR has proven a useful instrument for identifying and assessing a broad range of ethical issues in different research and development projects. However, it does not indicate how these issues should be handled with respect to the concrete design and implementation of system components. Therefore, up to this point it remains largely unclear to engineers and developers what exactly they should do, or not do, or how they should do it, during their day-to-day work creating the system. Another limitation of MEESTAR is the sole consideration of potentially negative aspects, because it is meant to safeguard against harm as "the minimum ethical requirement" (Manzeschke et al., 2015). Whenever the potential negative consequences of a system are within a tolerable range, they must be traded off against expected positive outcomes. The second part of AWOSE aims at tackling both of these shortcomings.

9.3 Integration with Worth-Centred Development

The classical user-centred design and usability engineering methodologies focus on properties of users and their interaction with a system, while user experience is mainly concerned with users' aesthetic and emotional perceptions before, during and after system usage (ISO 9241-210). Worth-centred development (WCD; Cockton, 2006) goes beyond these considerations and demands that the worth that is generated for people or organizations by (using or applying) a system should be targeted as the prime focus during development. Hereby "worth" may refer to any kind of individual or collective ethical, practical, financial, emotional, or other benefits and positive outcomes of system usage. The notion of worth even includes "unfelt needs" (Cockton, 2006), i.e. worth that is not yet consciously known to or explicable by potential stakeholders. This facilitates the creation of highly innovative products whose worth may only become evident by the time they are used. American psychologist Frederick Herzberg (1964) considered human motivation as an interaction of two main factors: Motivators, whose presence generates satisfaction (e.g. appreciation or professional success), and hygiene factors, whose absence creates dissatisfaction (e.g. payment or safety). With respect to this model, WCD asks system designers to focus *"on the* worthwhile*, that is, things that will be valued, as manifested in people's motivation, individually or collectively, to invest one or more of time, money, energy and commitment. [...] In short, worth is a motivator [and] designing worth means designing things that will motivate people to buy, learn, use or recommend an interactive product, and ideally most or all of these."* (Cockton, 2006, p. 168) In this sense, the focus of WCD seems to be contrary to that of MEESTAR at first glance. WCD's creator acknowledged that weighting of positive and negative aspects is required, such that the resulting system design "delivers sufficient value to its intended beneficiaries to outweigh costs of ownership and usage" (Cockton, 2008c, p. 60). Contrary to the precursory framework called "value-centred design" (Cockton, 2005), the existing publications on WCD do not include a well-defined process model but rather constitute a set of approaches to apply throughout development. Arguably, the most important and widely used of those approaches is the Worth Map, a specific type of

diagram that supports and structures systems design with respect to WCD's premises.

9.3.1 Worth Mapping basics

In market research, means-end chains describe the (expected) causal connections between product features, customers' emotions and their motivation for buying. WCD adapts this approach to the principle of "designing as connecting": Dependencies and connections between different system designs, usage, user experiences, stakeholders, and evaluation metrics are analyzed and expressed by connecting the respective elements (Cockton, 2009a, 2009b; Cockton, Kirk, Sellen, & Banks, 2009; Cockton, Kujala, Nurkka, & Hölttä, 2009), e.g. visually in a diagram consisting of boxes and arrows. In WCD and AWOSE, the elements of means-end chains can be materials and other components, features, qualities, and, finally, worth of a specific system. Different methods can be used to identify the elements of means-end chains with respect to a specific system:

Brainstorming about human needs, desires, aversions, motivations, habits and technical possibilities as well as experiences with comparable systems and current trends can be conducted by an interdisciplinary team (Cockton, 2008a).

Laddering is a technique originating in clinical psychology where it is used as an instrument to find out about the understanding that people have regarding their social relationships by asking them to describe people meaningful to their own life and then recursively querying about the meaning of constructs used in their description (Cockton, 2008c). This yields extensive information about people's personalities and values. In marketing, laddering is used to uncover the relation between personal values and the perceived benefit of products. To this end, customers are asked to name product attributes that are important to them. Afterwards they are recursively asked why these attributes are important, "repeating this ascent up the ladder until a consumer can only say that something really matters to them" (Cockton, 2008b, p. 293). The same principle can be applied in WCD to identify means-end chain elements and their association related to a system.

Sentence completion tests are semi-structured, projective surveys that have been applied e.g. as personality tests (Holaday, Smith, & Sherry, 2000), for determining managers' motivation (Brief, Aldag, & Chacko, 1977), and in consumer research (Donoghue, 2000). Participants are asked to finish given incomplete sentences according to their own first association. Such an incomplete sentence could be: *"Professionally, the most important thing for me is..."* It has been reported that sentence completion using incomplete sentences derived from a set of general and project-specific human values yielded better results than interviews and, despite the predetermined beginning of the sentences, is open enough not to subject participants to priming effects (Cockton, Kujala, et al., 2009). While brainstorming has the potential to generate any kind of means-end chain elements and laddering identifies complete means-end chains starting from system features upwards, sentence completion aims at uncovering people's most important values and motivators, i.e. system design goals in the form of intended worth. In AWOSE, any of the abovementioned methods, or a combination thereof, can be used to identify worth in the sense of motivators and desired positive outcomes of system usage. In the ADAMAAS research project, the intended worth was originally derived from the predefined project goals and extended through brainstorming in interdisciplinary groups and stakeholders surveys.

All identified means-end chains and possibly unconnected elements are then combined into a single diagram, the Worth Map of a system. Worth Maps are the core artifacts of WCD. They serve as a basis for discussion, to represent development goals and means for accomplishing them, and for planning evaluations. In order to create a Worth Map, the means-end chains are merged at common elements (if such exist) and complemented with isolated chains and elements. In AWOSE, the initial Worth Maps are iteratively refined and extended during development. Depending on the scope and complexity of the system, Worth Maps can become quite large and complex as well. It is therefore advisable to use software tools that facilitate making changes and adjustments to the diagrams with levity. Microsoft Office Visio has been recommended for this purpose (Cockton, Kujala, et al., 2009), with LibreOffice Draw constituting a serviceable open source alternative, and

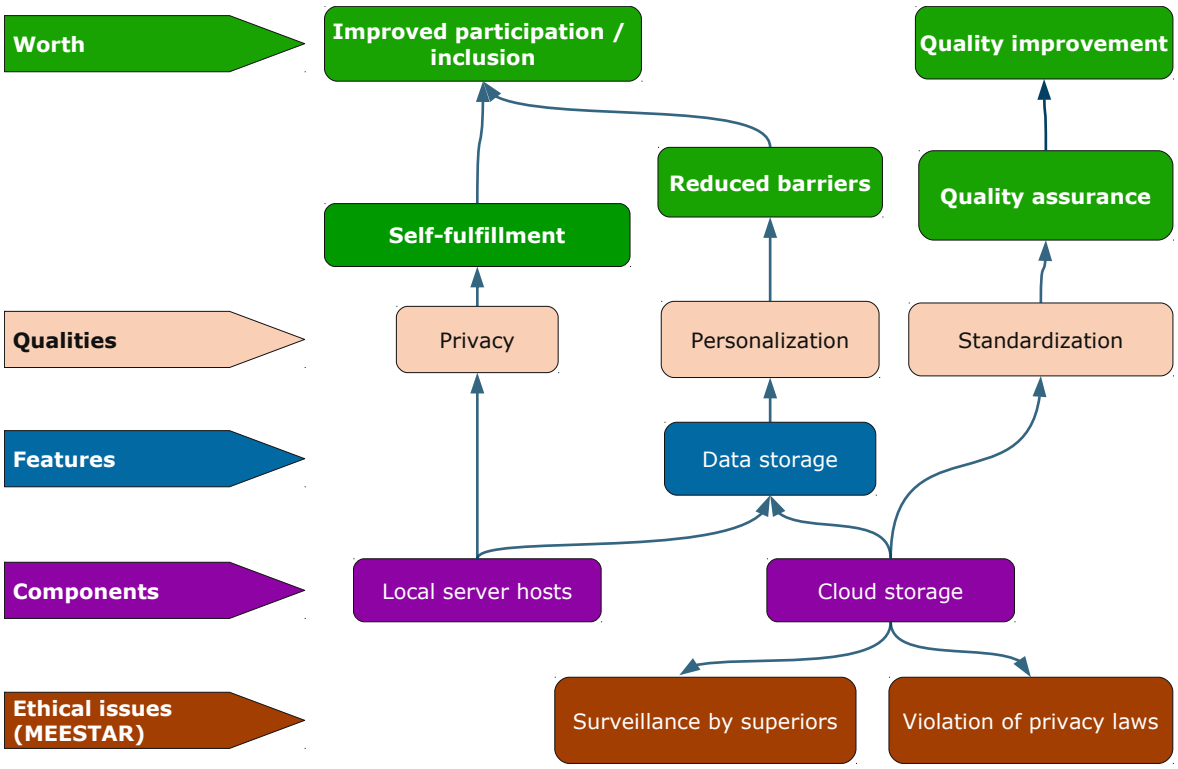

Figure 9.2: **Excerpt of a Worth Map sketch** from Project ADAMAAS representing a design decision regarding storage of data about users for the purpose of individualized adaptation. Purple boxes describe system components, blue boxes list possible features, orange boxes show qualities, and green boxes indicate worth, i.e. positive outcomes of the system. Red boxes in the bottom row represent ethical issues identified with MEESTAR.

a dedicated tool for Worth Mapping was developed in a computer science diploma thesis at Paderborn University (Strotmeier, 2001).

9.3.2 Integrating ethical issues in Worth Maps

It was explicitly not the original aim of WCD to avoid negative outcomes of a system's usage at all costs (Cockton, 2012), but rather it was to focus on worthwhile outcomes. Arguably, it depends on the assessed severity of ethical issues in how far they should prevent or restrict the usage of a system or its components. Additionally, when ethical issues apply only to some specific design variants or ways of implementing a given feature, often less critical alternatives can be chosen. This is highly important information for designers and engineers to keep in mind during development. The structure of Worth Maps is well suited to facilitate this. To this end, the Worth Maps

in AWOSE are extended by integrating the output from the first part based on MEESTAR, i.e. an additional layer of elements is added to the bottom of the Worth Map diagram, representing all ethical issues that have been identified (see bottom row of Figure 9.2). The workshop participants from the first part then come together again to investigate the associations between the Worth Map elements (e.g. system components and features) and ethical issues. These connections are graphically indicated with arrows in the Worth Map.

9.3.3 Increasing the expressivity of Worth Maps by UML integration

There are two basic types of connections between Worth Map elements, which indicate that positive or negative outcomes are either enabled (solid lines) or disabled (dashed lines), as described by Cockton, Kujala, et al. (2009). These simple types of connections are often insufficient for describing the interrelations of complex systems. Therefore, Worth Map diagrams in AWOSE use relationship notations from Unified Modeling Language (UML) structure diagrams when required, i.e. specific types of lines and arrows to indicate relations between elements such as implementation, dependency, or composition. This is especially useful for the layers containing technical descriptions (e.g. system components and features). With proper tool support, a broad range of UML diagrams can directly be integrated as elements or layers into Worth Maps. Zooming in and out of such extended Worth Maps as a project's master diagram can help increase designers' and developers' awareness of the "bigger picture", i.e. each system component's relation to features, qualities and, finally, desired positive outcomes of system usage, as well as ethical issues that should be considered.

9.3.4 Ethical and worth-related system evaluation

Designing and developing systems in a way that fulfills specified goals is a basic concept of engineering (Butler, 1985). Process models like the usability engineering lifecycle (Mayhew, 1999) and the human-centred design process from ISO 9241-210 require the definition of specific goals regarding usability and corresponding requirements as the basis for system evaluations. This

highlights the importance of usability and user experience as non-functional requirements within these frameworks. The definition of specific usability goals facilitates the planning of usability evaluations (Quesenbery, 2001) and may even guide the overall system design process by establishing the most important values and goals the resulting product shall fulfill (Quesenbery, 2003). In iterative design processes, usability metrics can be used to decide if further iterations are necessary (Whitefield, Wilson, & Dowell, 1991). Leading IT companies like IBM, Microsoft, and Google use the evaluation results as a formal basis for deciding upon product release (Beauregard & Corriveau, 2007). The consensus is that the definition of specific goals and corresponding evaluation is beneficial. However, different approaches have been taken in deciding how to define suitable evaluation criteria. In the practice of usability engineering, evaluators often resort to generic measures that are easily operationalized, e.g. the time users require for task completion, or the number of errors they make when interacting with the system, independent of the actual relevance of these measures in a given context. Instead, project-specific proprietary measures should be defined with respect to each system, its context of use, and the overall goals (Beauregard & Corriveau, 2007; Cockton, 2008b).

A central premise of WCD and AWOSE is that usability does not carry inherent worth, but rather that it is often a necessary means to superordinate ends. In AWOSE, the definition of metrics for system evaluations works as follows: On the one hand, the Worth Map elements are analyzed starting at the top level (worth) and then possibly going down to lower levels of system qualities, features, or components, if and only if the associated higher-level elements cannot be measured. As an example, imagine an assistance system that is supposed to help people with handicaps to learn how to perform working steps more quickly and independently from their teachers. The faster learning rate and independence may be considered positive outcomes or worth. Therefore, if it was possible to measure the users' degree of success in learning the working task and their independence before and after introducing the assistance system, the generated worth could be assessed without doing classic usability tests with the system. However, it may be that participants cannot be exposed to a system prototype of unknown quality, because it might

confuse and irritate them, or a proper assessment of their success at learning new working tasks would require too much time and resources. In this case, evaluation may need to resort to lower-level metrics like the understandability of the wording and quality of icon design of the assistance system's user interface. On the other hand, AWOSE also requires system evaluators to consider potential ethical concerns. Most of these, and coercively those with critical severity ratings, should not make their way into system implementations in the first place. For the remaining issues, evaluation metrics must be defined. For example, imagine again the abovementioned assistance system. An ethical issue might be that users succeed well in performing the working tasks when using the system but rely heavily on its assistance instead of learning the task on their own. This would indicate that an unwanted dependence on technology has been induced, which would be contrary to the goal of facilitating learning processes. A corresponding evaluation metric for this issue could be to regularly assess and compare users' task performance with and without the assistance system.

Now that the ingredients and rationale of AWOSE have been outlined, the next section proposes a process model to structure agile development endeavors with respect to ethical and worth-related aspects.

9.4 An agile process model

While the previous proposals for integrating human-centred aspects and agile methods mainly focused on conventional usability and user experience (see e.g. Holzinger et al., 2005; Lee et al., 2009; Memmel et al., 2007; Obendorf & Finck, 2008; Singh, 2008), AWOSE establishes worth-related and ethical aspects as the primary concern.

The AWOSE process model (Figure 9.3) assigns responsibilities to three different roles:

- The *customer* plays a similar role as in the Extreme Programming (XP) (Beck, 2000) and can be related to Scrum's concept of a "product owner". He or she should be available for the system development team during the whole project. However, the role can be assumed by differ-

ent individuals over time, with the additional benefit of having different people as test users for "quick-and-dirty" formative evaluations.

- The *worth designer* fulfills comparable duties as usability engineers, interaction designers, or user experience specialists, but aims to support the worth- and ethics-oriented goals.
- The *developer* is responsible for technical planning and system implementation.

Note that in large projects usually a handful of people will share the roles of worth designers and developers. The (non-iterative) preproduction phase of AWOSE starts with the definition of a "system vision" by the customer. As discussed in the context of AWOSE's first part, the granularity of the resulting description can vary and depends on project properties. The worth designer then organizes an interdisciplinary workshop to identify and assess ethical issues based on MEESTAR, as well as the intended worth of the system. Next, the worth designer conducts context of use analyses (as in other human-centred design processes), producing Persona stakeholder models and other artifacts. In parallel, the developer engages in a technical exploration phase as in XP in order to find out if, how and with what expected effort specific requirements can be fulfilled and features implemented. Subsequently, the results are consolidated into an initial Worth Map. On this basis the customer and worth designer define superordinate project goals and evaluation metrics, and the customer decides upon a set of features that shall be implemented in the first iteration.

After that, the iterative production phase starts (light-blue ellipse in Figure 9.3). The duration of iterations depends on the properties of individual projects and should be kept constant during development. The developer implements a set of features that have previously been defined by the customer and designed and tested by the worth designer. In parallel, the worth designer creates prototypes for a set of features that are supposed to be implemented in the subsequent iteration, and conducts formative evaluations with them. Finally, the three roles get together for a meeting in which the developer reports on the expected costs for implementing pending features, the worth designer updates the Worth Map with relevant new information that has been gathered

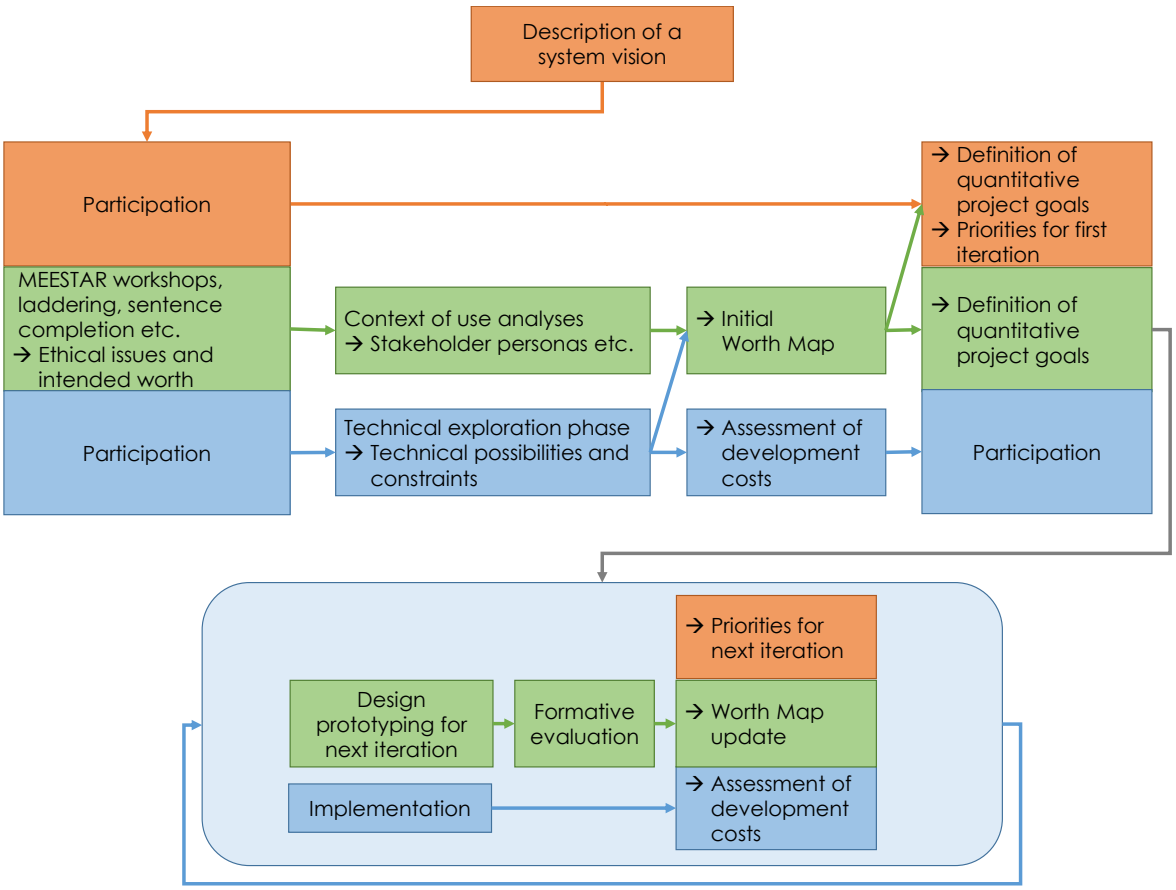

Figure 9.3: **The AWOSE process model** with responsibilities of the customer or product owner (orange), worth designers (green) and developers (blue).

in the meantime, and the customer then decides upon the features that shall be designed in the next iteration. An important characteristic of AWOSE is that both the selection of features and choices among alternative ways of implementing these should primarily be based on the current Worth Map by evaluating the ethical assessment and expected generation of worth that are associated with each of the still outstanding features. After this meeting, the next iteration starts.

9.5 Discussion

Adequate consideration of ethical aspects in research and development must balance diligence and practical feasibility. As Zhu and Jesiek (2017, p. 677) noted, *"preferable ethical decisions are "workable", i.e., they need to be both*

ethically justifiable and practically plausible". Whenever an agile approach to development can be adapted, the AWOSE methodology may help structure the process and suggest how to apply effective methods to satisfy these requirements. The two main parts of the methodology, MEESTAR and WCD, complement each other regarding their goals and the insights generated by their application. MEESTAR aims at safeguarding against harm by identifying ethical sensitivities related to a system (Manzeschke et al., 2015). A broad range of projects has demonstrated its applicability, including and beyond age-appropriate assistance systems such as ADAMAAS. The second component of AWOSE, WCD, aims at designing systems that deliver value in the world, which endures after interaction (Cockton, 2006), and established the use of worth map diagrams to support this goal during system design. In addition to providing a systematic structure for effective integration of MEESTAR and WCD, AWOSE incorporates several improvements on its heritage. The MEESTAR set of ethical dimensions was extended to include nature-related aspects, such as sustainability of systems, and the important step of determining ethical severity ratings has now been procedurally defined. Worth maps have, as a result of several iterative optimizations, already been described as "state of the art in values focused methods" (Cockton, 2012, p. 4). Nonetheless, the fusion with UML diagrams as proposed in this chapter makes them potentially more useful for complex development projects and may popularize their use in organizations with a strong technocratic orientation. On a final note, using these improved worth maps the AWOSE methodology may also support large-scale system-level analyses as demanded for example by Borenstein, Herkert, and Miller (2017) in the context of autonomous driving.

Compared with many other approaches to assessing ethical issues in engineering (e.g. Alkhatib & Abdou, 2017; Hofmann, 2017; Hofmann, Haustein, & Landeweerd, 2017; Kermisch & Depaus, 2018; Lokhorst, 2018), the approach taken in AWOSE has several conceptual benefits. It is comparatively "open" in the sense that, albeit referring to a set of high-level ethical dimensions in order to stimulate and structure the brainstorming process, it does not impose a specific predefined list of questions that may unduly bias and distort results. Furthermore, it supports the assessment of society-level concerns related to public accountability of research as called for by Leese

(2017), and it is embedded in an overarching process model. On a theoretical level, AWOSE differs from VSD in that it explicitly distinguishes between avoiding ethical issues on the one hand and creating worth on the other hand, according to Herzberg's two-factor theory. Hereby "not creating worth" (i.e. not increasing stakeholder's motivation for system use) does not necessarily imply ethical issues, whereas neglecting "hygiene factors" might constitute an ethical issue. Other than VSD's definition of "values", AWOSE's "worth" does not need to be something that people "consider important in life" (Friedman et al., 2008, p. 70) but only something that motivates them to use the system. The VSD approach (Davis & Nathan, 2015, p. 22) requires that "designers *must* attend to values supported by theories of right, which are obligatory, and *may* attend to values supported by theories of the good, which are discretionary". The terminology and methodological approach of AWOSE makes a conceptually related but more explicit distinction by requiring that (negative) ethical issues *must* be prevented or mitigated, while any kind of (positive) worth *may* be created through a system in order to motivate its usage. For example, "looking hip and stylish" would probably not be considered "important in life" by many people, but nevertheless it may be a factor motivating them to buy and use such things over less aesthetically pleasing alternatives. In comparison with Spiekermann's E-SDLC approach, AWOSE embraces agile principles more genuinely. E-SDLC supposes that the prioritization of values is finished before iterative software engineering even starts. In AWOSE's production phase, worth maps are updated at the end of each iteration. This enables the "customer" to establish priorities for the next iteration with explicit reference to the intended worth and ethical issues, even when the requirements and system architecture have changed arbitrarily since the previous planning meeting. Apart from that, the two methodologies share many conceptual similarities. For example, both approaches refer to Personas as stakeholder models for ethical analyses. Notably, Spiekermann only presents the "Ad-Hoc Persona" variant (as Norman called it), whereas AWOSE prefers data-driven stakeholder models whenever possible. The decomposition or conceptualization of values as described by Spiekermann (i.e. breaking a value "down into the subdimensions that constitute its essence", p. 205) maps directly to means-end chains of worth elements in AWOSE's

worth maps. Furthermore, AWOSE and E-SDLC (as well as VSD for that matter) concordantly promote choosing design alternatives with respect to their value-/worth-related impact, albeit at different points in time within the process.

As of now, practical experience with the AWOSE methodology is limited to Project ADAMAAS. Approaches from AWOSE have been applied in this project to analyze ethical aspects and guide development of the smart glasses assistance system with promising results and to its stakeholders' satisfaction. However, it remains to be seen how the methodology scales and adapts to larger developments projects and other types of research. From a conceptual and theoretical perspective, AWOSE using MEESTAR's default dimensions (plus nature-related considerations) certainly suits the development of ICT-based assistance systems best, but it should be applicable to other technical systems and potentially other forms of engineering as well, as long as these allow for rapid prototyping and short iterations during development. In this case, the ethical dimensions and the integration of technical descriptions in worth maps may need to be aligned accordingly.

People Propose,
Science Studies,
Technology Conforms

– Don Norman (2014)

Chapter 10.

User-centered engineering activities

A small fraction ($\approx 20\%$) *of this chapter is based on lines of argument from my German-language Bachelor's thesis in computer science.*

The AWOSE methodology from the previous chapter considers usability-related system features as potential means to higher-level ends. In Project ADAMAAS, it was evident that usability aspects would play a fundamental role in bringing the potential benefits of the system to fruition and unlocking its intended worth. The following sections recapitulate the selection of appropriate user-centered design methods, stakeholder modeling activities, and system usability evaluations that were carried out during the project.

10.1 The usability method selection tool

Since the early 1990s, scientific research in the field of human-machine interaction has flourished, and reams of more or less different methods have been designed to support the development of usable systems. However, Furniss, Blandford, and Curzon (2007) pointed out that many of these methods are never carried over from the academic world into practical application by usability experts. One reason for this may be that it is difficult even for experts to keep track of all the methods that have been developed by scientists and practitioners around the world. This problem was already mentioned by Olson and Moran in 1995, although the amount of existing methods at that time was certainly much more manageable than today. Furthermore, there

have been a large number of different suggestions and recommendations regarding the selection of suitable methods, which offer no uniform concept and contradict each other to some extent (see Bevan, 2003). In 2002, the technical report ISO/TR 16982 offered in-depth information and guidance on twelve popular human-centred design methods and categories of methods including their advantages and disadvantages, as well as a concrete concept for evaluating the suitability of these methods based on a broad range of criteria. The 18 criteria were divided into six categories:

- The phase of the (software) life cycle,
- project environment constraints,
- user characteristics,
- task characteristics,
- properties of the product or system to be developed, and
- the extent of in-house expertise in the area of ergonomics.

ISO/TR 16982 included tables that assigned ratings to each method with respect to each of these criteria on a five-point scale ranging from *"not applicable (N/A)"* to *"recommended (++)"*. In order to determine the most suitable methods for a particular project, the report recommends a preselection based on the phases of the life cycle and the project characteristics. The selection of positively evaluated methods should then be further refined based on the remaining criteria. Unfortunately, the tables in ISO/TR 16982 do not enable an efficient review and selection of suitable methods, because the ordinal-scaled symbolic ratings cannot easily be condensed to an overall judgment. Therefore, Fischer, Strenge, and Nebe (2013) suggested the following mapping:

recommended (++)	$\mapsto 1$
appropriate (+)	$\mapsto 0.75$
neutral	$\mapsto 0.5$
not recommended (-)	$\mapsto 0$
not applicable (N/A)	$\mapsto -\infty$

	Observation of users	Performance-related measurements	Critical Incidents analysis	Questionnaires	Interviews	Thinking aloud
Development - Requirements analysis	0.86	0.75	NA	0.84	0.87	0.69
Development - Architectural design	0.83	0.78	NA	0.81	0.84	0.69
Development - Qualification testing	0.83	0.78	NA	0.84	0.87	0.67

	Collaborative design and evaluation	Creativity methods	Document-based methods	Model-based methods	Expert evaluation	Automated evaluation
Development - Requirements analysis	0.84	0.72	0.70	0.59	0.70	0.54
Development - Architectural design	0.84	0.75	0.73	0.59	0.70	0.57
Development - Qualification testing	0.84	0.70	0.70	0.59	0.70	0.57

Figure 10.1: **Project-specific suitability ratings of different usability methods.**

It should also be possible to assess the 18 criteria of ISO/TR 16982 not only via dichotomous decisions (i.e. fully confirm or reject) but also in a more fine-grained way. Representing these decisions as real values then allows for calculating the weighted average as an overall suitability rating for each method. In Project ADAMAAS, this concept was implemented as a spreadsheet tool, which enabled automatized calculation of suitability ratings. The application of 17 criteria from ISO/TR 16982 (excluding the life cycle phase) has been assessed independently by five project partners. The median value of these assessments then served for the calculation of method suitability ratings, which are shown in Figure 10.1. Unsurprisingly, the special characteristics of the project's targeted users and other project properties resulted mainly in recommendations for collaborative methods with extensive direct involvement of users rather than e.g. automatized usability evaluations. The subsequent user-centered design activities in the project were hence strongly oriented in that direction. Important groundwork for identifying and involving the right people as prospective (test) users and other relevant stakeholders

was a proper data-driven modeling of their characteristics, which also complies with the AWOSE methodology (see Section 9.2.2).

10.2 Data-driven stakeholder modeling

The targeted stakeholders of the ADAMAAS system were associated to three different predefined contexts of use:

- A sheltered workshop at Bethel proWerk, a diaconal organization that helps people with mental disorders and disabilities,
- the retirement home Bethel Breipohls Hof,
- and the company Hettich, a leading manufacturer of furniture fittings that was introduced in Chapter 7.

The ADAMAAS system was meant to assist different activities related to education, daily life, and professional life in these contexts of use: At proWerk, learning to build wooden birdhouses; at Breipohls Hof, operating a high-tech automatic coffee machine with a touch interface; and at Hettich, manually assembling product mockups. In the end, two to three personas for each of these application scenarios represented the most relevant stakeholders in terms of the following characteristics:

- Age,
- role within the context of use,
- special physical and mental properties, such as disorders or disabilities,
- and goal- or feature-related requirements.

The amount of data on stakeholder properties differed notably between the three scenarios due to project-related constraints and prioritization. Therefore, different procedures were applied to transform the raw data properly into persona descriptions.

Most data was available for the proWerk scenario. Based on previous methodical propositions for data-driven persona modeling (Brickey, Walczak, & Burgess, 2012; Sinha, 2003), an idiosyncratic combination of methods from

exploratory data analysis based on principal components analysis and descriptive statistics was used to first identify independent sets of correlating requirements and then typical properties of the corresponding stakeholders. The algorithmic delineation in Appendix A outlines this experimental new approach. For example, the mean age of those stakeholders that had the highest scores on one of the principal components and low scores on other principal components was calculated. Then a persona representing this subgroup was assigned its mean age alongside a memorable name that used to be popular in the corresponding hypothetical year of birth. This comparatively rigorous data-driven approach to persona creation can be seen as an extension and refinement of the method proposed by Sinha (2003). Theoretically, the resulting set of persona descriptions could have been evaluated by gathering data about the multivariate prevalence rates of their attributes among the real stakeholder population. However, this is practically infeasible for sufficiently detailed personas due to the "curse of dimensionality"[1] (Chapman & Milham, 2006). As a pragmatic alternative, domain experts were informally asked to judge the degree to which the persona set appeared internally consistent and covered relevant stakeholder properties. Universally positive feedback indicated that the approach may have been suitable. However, no definite conclusions can be drawn in that regard yet due to the highly informal nature of this expert assessment and lack of comparison with other approaches.

Fewer data was available on Breipohls Hof stakeholders, so persona creation was based on simple descriptive statistics (i.e. measures of central tendencies) for characterizing one representative stakeholder model for each of the two relevant roles (inhabitants and staff members). The inhabitant persona was established as the primary persona for this scenario. When creating the secondary persona representing the staff members, the specific requirements of the first persona were disregarded in order to increase the distinctness and memorability of the persona set.

The least amount of explicit data about stakeholders was available for the scenario at Hettich. Stakeholder modeling on the academic side therefore resorted to ad-hoc persona creation based on limited observations of workers

[1]The number of possible feature combinations grows exponentially with the number of features, which is why more detailed personas inevitably have lower prevalence rates.

and context of use, as well as subjective appraisal. Hettich's human resources department and the responsible ADAMAAS project contacts then evaluated these initial persona drafts, which were finally adjusted based on that feedback.

During the design and development of the ADAMAAS system prototype, the personas served not only as a reference for identifying and assessing ethical issues and intended worth, as demanded by the AWOSE methodology, but also as a guideline for prioritizing system features and selecting stakeholders for formative evaluations. Towards the end of Project ADAMAAS, the iterative cycle of evaluation and adjustment was terminated by the evaluation activities reported consequently.

10.3 Evaluations

The concluding evaluation activities in Project ADAMAAS were organized in a bipartite way: Verifying that a satisfactory level of usability had been achieved and defining worth-related measures for long-term assessments.

10.3.1 Usability tests

At Hettich, a research prototype of the ADAMAAS 3D in-situ AR system for manual assembly assistance of a drawer mockup was tested. The accuracy of the SDA-M-based cognitive prediction component had already been evaluated separately before (see Chapter 7). The computer vision action recognition component was still in development and therefore had to be tested separately later. Thus, this evaluation focused on the design of the AR component, which could show assisting instructions for each assembly step at the designated place of action. To this end, an ideal prediction of individual problems and perfect action recognition were simulated with the *Wizard of Oz* technique: Whenever participants attempted to make a mistake or did not know how to proceed, an experimenter triggered the required AR instruction for the current action. The participant sample of $N = 28$ employees included professional carpenters and joiners who had extensive experience in manual assembly, as well as clerks and office workers with limited task-related experience. They were between 21 and 70 years old with a mean age of 41.8

years ($SD = 11.2$). The majority (75%) of them were male. First, participants were shown an instruction video that demonstrated all assembly steps. Next, they were asked to put on the Microsoft HoloLens AR smartglasses running the ADAMAAS software and assemble the drawer system mockup. Whenever they attempted to execute a wrong action, a 3D in-situ instruction for the current step was shown. After assembly the System Usability Scale (SUS) of Brooke (1996) was administered. The SUS yielded a total score of 73.2, which is interpreted as a "good" (Bangor, Kortum, & Miller, 2009) and above-average result (Sauro, 2011). The learnability factor, as defined by Lewis and Sauro (2009), reached a score of 76.8 points.

Three female inhabitants of the Bethel Breipohls Hof retirement home complex with a mean age of 86.0 years visited a lab at Bielefeld University's CITEC to test the final version of the ADAMAAS assistance system for using a high-end automatic coffee machine with touchscreen. In the test task, the water container of the coffee machine was empty and had to be refilled before coffee could be made. An analysis of participants' task-related mental representation structures with the SDA-M-based CASPA algorithm confirmed the assumption that none of them had any useful previous knowledge about interacting with this specific machine. This was reflected without exception by negligible estimated chances of correct action selection. Appendix B specifies the textual descriptions of actions and the intended sequence that have been used for this test. Therefore, the ADAMAAS system assisted each and every step. The SUS score of 56.7 points fell in the range between "OK" and "Good". Interestingly, participants' overall verbal judgments of the ADAMAAS system features in a subsequent interview were universally positive, but they made unfavorable remarks concerning properties of the coffee machine and the HoloLens hardware. This suggests that the mediocre SUS score in this scenario might be ascribable to some extent to halo effects.

The AR assistance for birdhouse assembly was tested at Bethel proWerk with $N = 6$ test users that had mild mental disorders or psychological issues. Participants were between 18 and 38 years old with a mean age of 22.5 years ($SD = 7.7$). All but one of them were male. This test was primarily intended as an assessment of the visual AR assistance design concepts' generalizability to special user groups. A SUS score of 85.8 was reached, which is interpreted

as an "excellent" result (Bangor et al., 2009). Qualitative user statements were gathered as well, which turned out universally positive, including the system was *"cool and fun to use"*, *"good and helpful"*, and *"overall great"*.

It should be stressed that, due to time limits and other project constraints, these usability tests had rather small sample sizes and used research prototypes of ADAMAAS system components that were not mature enough for productive use. The results should therefore be interpreted with caution, since long-term effects and influences of different usage contexts have not been explored yet.

10.3.2 Worth measurement

Complementary to conventional usability evaluations some extensive groundwork was carried out in Project ADAMAAS with regard to long-term assessments of actual worth generated by the prototyped system once it would become market-ready. Interdisciplinary cooperation between academic and application partners resulted in a set of high-level worth elements (left) and associated measures (right):

efficiency	$\mapsto$ manufacturing time, scrap rate, training period
product quality	$\mapsto$ outcomes of quality assurance
user acceptance	$\mapsto$ dropout rate
safety	$\mapsto$ number of incidents, subjective safety (questionnaire)
data security	$\mapsto$ expert review
participation	$\mapsto$ increase of user group size
self-dependence	$\mapsto$ frequency of call button usage, extent of supervision

These measures were contrived specifically for AR-based cognitive assistance in the contexts of use that were targeted in Project ADAMAAS, but clearly many of these measures apply to a broad range of assistance systems. Anyhow, the main point of this list is to serve as an illustration of how to translate fuzzy definitions of high-level worth into specific, quantifiable measures. Hopefully, this paves the way for more meaningful evaluations of future cognitive assistance systems' impact and worth in the long run.

Chapter 11.

General discussion

This original chapter concludes the thesis by discussing its scientific accomplishments and provides an outlook on future research possibilities.

The primary goals of the scientific endeavors reported in this thesis have been to automatize the analysis of data about individual task-related mental representation structures, to evaluate the usefulness of these automated approaches, and to investigate how they could be integrated into technical systems for cognitive assistance. By pursuing these goals, this thesis explored one of countless conceivable approaches to human error prediction and expertise assessment. Simultaneously, it investigated one of various conceivable ways to tackle the issue of user-adaptive adjustment of assistance systems, e.g. those using AR to superimpose users' field of view with virtual elements to provide helpful instructions, depending on users' individual cognitive properties. The following sections provide an aggregated overview about the key findings and limitations concerning these goals, reflect on the nature of this research and its accomplishments, and suggest directions for prospective investigations.

11.1 Key results

With respect to the research questions that have been defined in Section 2.4, the evidence acquired in this work suggests the following answers:

RQ1: *Can the procedures for analyzing task-related mental representation structures based on SDA-M be further automatized with algorithmic approaches?*

→ **Yes**, at least for a certain set of applications, as has been demonstrated in Chapter 4 by developing the AMPA and CASPA algorithms. These algorithms replace the previously required analysis step that involved human experts using SDA-M data visualizations (dendrograms). Some manual effort is still required to use the SDA-M method though. First, it is necessary that investigators and domain experts create textual and/or pictorial descriptions of the basic actions related to an activity once. An important limitation of the algorithmic approaches is that they require the activity to be represented by a sequence of distinct actions that do not overlap in time, which is not necessarily required in "traditional" SDA-M. Second, each study participant (or system user) must perform the split procedure to provide the raw data about his or her current mental representation structure related to the given activity. The latter may need to be repeated from time to time in order to update the data.

RQ2: *Do these "algorithmic SDA-M" analyses conform to the gold standard of "traditional SDA-M" that involves human expert assessments?*

→ **Yes, to a large extent**. Chapter 6 reported on a study that compared the algorithmic assessments from AMPA and CASPA with expert assessments based on SDA-M dendrograms empirically. In 84% to 86% of the test cases, algorithmic and experts' assessments matched. It was further found that human experts varied from one another in their assessments to some degree. Descriptively, the CASPA algorithm outputs correlated even higher with the mean experts assessments than individual human experts did on average. Therefore, algorithmic assessments can be considered at least equivalent to the traditional approach.

RQ3: *How well can algorithmic SDA-M analyses predict human errors related to different kinds of practical applications?*

→ **That depends** on the specific application and other factors, including the degree to which the theoretical assumptions of the algorithmic approaches are violated (see Section 4.1 and Chapter 7). The assumptions of currentness, context-independence, and most notably completeness are expected to be violated to some degree in most realistic practical applications. In the application-oriented empiric studies regarding two manual assembly tasks and a movement sequence, the CASPA$_d$ algorithm correctly predicted 68% to 74% of all errors, and CASPA$_i$ correctly predicted 72% to 77% of all errors. In addition to these studies, I supervised a Master's thesis and an international research internship project, which applied these approaches for research in real-time strategy (e-sports) and chess gambit action sequences. The overall accuracy values of algorithmic SDA-M analyses were significantly above chance level in all empiric studies. However, clearly some types of errors generally cannot be anticipated based on static SDA-M data, e.g. slips due to temporary distractions caused by a current external context.

RQ4: *Are algorithmic SDA-M analyses sensitive to changes in memory formation?*

→ **Yes**. This was investigated in a manual assembly task (see Section 7.3.2). Before participants went through a learning phase, CASPA estimated a low average probability of less than 19% that they would have chosen correct actions for building the designated construction. After the learning phase, CASPA estimated a significantly ($p < 0.0001$) higher average probability of correct action selection of almost 58%.

RQ5: *Are algorithmic SDA-M analyses applicable irrespective of skill levels, i.e. equally suitable for experts and laypersons in a particular domain?*

→ **Presumably yes**. The accuracies of algorithmic SDA-M predictions for experienced workers and laypersons were compared in an assembly study in industry (see Section 7.3.2). No differences have been found. However,

in general this may depend on the specific task, the constitution of expertise that is relevant for the task, and how it is reflected in the mental representation structures that can be measured by SDA-M. Further research in this direction might be needed. For example, in some domains experts on very high skill levels may not commonly make errors in terms of wrong basic action selections anymore but exclusively distinguish themselves from one another in terms of how efficiently or precisely they execute these actions.

RQ6: *How well can algorithmic SDA-M analyses assess people's formal expertise and overall performance in an activity compared to traditional SDA-M-based measures?*

→ **Very well**, as demonstrated in a karate movement sequence study. CASPA$_m$ is a newly conceived overall measure of task-related competency based on algorithmic SDA-M analysis (see Chapter 8). It correlated strongly and highly significant with the karate practitioners' formal expertise ranks (*kyu* and *dan* belt grades), as well as with their actual performances. CASPA$_m$ values even showed higher correlations with actual performance than formal expertise ranks did. Descriptively, CASPA$_m$ also showed stronger correlations with performance and expertise than traditionally used SDA-M-based metrics (λ and ARI).

RQ7: *Which methodological procedures should be used to take ethical issues and other system stakeholder requirements properly into consideration when developing cognitive assistance systems?*

→ The **AWOSE methodology** defines such procedures, e.g. to incorporate an algorithmic SDA-M component into an assistance system based on AR smart glasses. It has been described extensively in Chapter 9.

An additional, overarching research question that came up while working on the issues above, is which of the different computational approaches for automatizing SDA-M analyses would prevail in practice: The rather simple, static "winner-takes-all" approach of AMPA, or the more elaborated action selection mechanism of CASPA?

→ The results in this regard are not quite clear-cut, since the different al-

gorithms came out roughly comparable regarding their overall performance. The empiric investigations generally tended to indicate some advantages of the more sophisticated CASPA algorithm variants over the simpler AMPA, but so far no definite conclusions can be drawn in this regard based on the currently accumulated evidence.

11.2 Retrospection and reception

Any sincere appraisal of these results would first and foremost need to acknowledge the intended placement of this work in cognitive science, including related disciplines such as psychology and applications in sport science, as well as in human-machine interaction research. This interdisciplinary placement resulted in a method mix from basic and applied, empiric and speculative, rigorous and pragmatic research approaches, which aimed at generating the greatest possible benefit in terms of useful and reliable insights for developing appropriate means to providing cognitive assistance for human activities. For these purposes, I devised the algorithmic approaches (AMPA and CASPA) based on different theoretical assumptions and computational cognitive architectures and, together with my colleague Oleg Strogan, implemented them into the *QSplit SDA-M Suite* software tool. Subsequently, the new invention was empirically examined in various studies and for different purposes. Finally, a methodological concept has been established for integrating it as a "cognitive component" into user-adaptive assistance systems like ADAMAAS in an auspicious way, i.e. one that considers as many success-critical aspects as possible (including usability-related and ethical issues). In other words, the research concept of this work strove to match, integrate, fulfill and satisfy both the requirements for a scientific doctorate and the specific application-oriented goal definitions of Project ADAMAAS.

From a cognitive science perspective, the primary contributions of this thesis could be described as the creation of individual computational cognitive models for human action sequence execution based on, or derived from, different cognitive architectures and individual survey data. Experimental and observational studies investigated the empirical relationships between these models, as well as measuring units derived from them, and behavioral mea-

sures in different contexts. As Sun (2009) argued, this kind of research can likewise be considered as the definition and testing of cognitive theories.

From an application-oriented human-machine interaction research perspective, the work could be characterized as the engineering and user-centered evaluation of models, algorithms, (software-based) tools, and implementation-related proceedings to constitute an anticipatory cognitive component for technical systems, which determines the need for assistance in human action sequences.

In terms of public reception, Project ADAMAAS was undoubtedly successful. In 2018, the initiative "Land of Ideas", founded by the German government and the Federation of German Industries, praised it as an innovative and forward-looking project that benefits the country and its inhabitants, and honored it as one of 100 so-called "Landmarks in the Land of Ideas" (out of approximately 1,500 applicants). As a result, Project ADAMAAS contributed one of the two award plaques of this type that now adorn the entrance of the CITEC research building. Furthermore, the project was featured in numerous far-reaching news websites and German public mainstream media (e.g. the WDR 5 science broadcast *Quarks - Wissenschaft und mehr*, as well as the 1LIVE main news). It should be noted again though that the work reported in this thesis was only one of several parts of Project ADAMAAS that all contributed to its success. These other parts included research on AR attention guiding techniques (Renner, Blattgerste, & Pfeiffer, 2018; Renner & Pfeiffer, 2017; Renner & Pfeiffer, 2017a, 2017b), AR instruction design (Blattgerste, Renner, Strenge, & Pfeiffer, 2018; Blattgerste et al., 2017), eye tracking (Blattgerste, Renner, & Pfeiffer, 2018; Essig et al., 2016; Renner & Pfeiffer, 2017c), as well as object and action recognition by computer vision based on machine learning techniques (Schröder & Ritter, 2017a, 2017b). Providing appraisal from a purely scientific, application-agnostic perspective paradoxically seems inherently much more elusive. Despite ongoing efforts to establish a set of basic requirements for all serious scientific activities, such as rigorous peer reviews, ethics committee approvals, open access, public data availability, and conflict-of-interest statements, there is hardly any consensus yet regarding suitable criteria for assessing the "quality" of research. In practice, the scientific community usually resorts to surrogate

indicators such as the impact or number of publications. In this regard, Project ADAMAAS arguably succeeded reasonably well, too, with an output of at least 19 peer-reviewed research publications, 11 of which I directly contributed to as a (co-)author.

To end the retrospection of this research work on a philosophical note, it seems worthwhile to revisit a line of thought from Strenge, Vogel, and Schack (2019) by discussing under which conditions an assistance system like ADAMAAS may be considered an *anticipatory system* according to Rosen (2012) when it incorporates a prediction module based on algorithmic SDA-M analyses of its user's mental representation structures. Following Rosen's pertinent definition, this would be the case if the human user of the system was regarded as (part of) the system's "environment" and the system's internal predictive model *"provides an alternate description of the entailment structure of the mapping representing the [biological] process itself"* (Louie, 2010). Importantly, this definition distinguishes *models*, which actually describe physical or biological processes and require an understanding of those, from simpler *simulations*, which merely describe the effects of a process. In this sense, statistical "models" like curve-fitting would only qualify as simulations but not as models. Since the present work was concerned with the cognitive processes of (correct or incorrect) action selection, this requirement seems to be satisfied if (and only if) the predictive model was grounded on neurocognitive actualities. Arguably, both Schack's CAA-A theory underlying the SDA-M method and Anderson's ACT-R theory, from which CASPA's calculations are derived, may be considered as sufficiently well-grounded in this regard. Furthermore, the predictive model M of an anticipatory system S should be *"equipped with a set E of effectors that operate either on S itself or on the environmental inputs to S, in such a way as to change the dynamical properties of S"* (Louie, 2010, p. 26). Such effectors could for example be the visual or auditory displays of an assistance system, which cause its user to behave in a different way, i.e. *"the effect of the model M creates a discrepancy – S would have behaved differently if M were absent"* (Louie, 2010, p. 26). According to Louie (2010, p. 28), such a "predictive or anticipatory mode" would cause a system to *"become more like an organism, and less like a machine"*.

11.3 Outlook

The research presented in this work shed light on some of the most fundamental and essential issues concerning computational analysis of task-related mental representation structures and its applications for user-adaptive cognitive assistance, but it also raised a number of further questions, which constitute interesting opportunities for future investigations:

- Cognitive assistance systems should not only enable proficient users to effortlessly discard unneeded suggestions in case of falsely predicted errors but also allow them to ask for help whenever the system did not anticipate they would have an issue in the current situation. It might be useful to establish a general standard for these operations in terms of user interaction design.
- The assumption of context-independence for generating error predictions with AMPA and CASPA is obviously merely a theoretical one. With the possible exceptions of tightly controlled and highly artificial VR or laboratory settings, real contexts of use in private or professional applications should be expected to involve a broad range of external factors, which influences the probabilities of errors in human action sequences substantially. How can these environmental factors and the resulting current, transient user state be measured and considered in error prediction and user adaptation components? A possible direction for further research might be to integrate electroencephalography or other physiological measurements (e.g. breathing rate, electrodermal activity, heart-rate variability, or pupil dilation) to assess relevant user states and adjust the system accordingly.
- Strenge et al. (2019) speculated that a user-adaptive cognitive assistance system might help specific target groups overcoming insecurity and hesitation to tackle unfamiliar activities. Is this actually the case, and if so, under which conditions?
- How often should users repeat the SDA-M split procedure to update the data? This directly depends on how stable or volatile the mental representation structures for a given task are. Future investigations could strive to

estimate a proper updating frequency as a function of the intensity of deliberate practice and other activities related to learning processes a specific person engages in. It might also be possible to find a proper way to extrapolate users' individual learning curves in order to minimize the frequency of split procedure updates.

- As discussed in Chapter 8, an extensive research and development project could build an assistance system and empirically test its impact on the quality and efficacy of deliberate practice compared to unassisted exercise and to training with a human coach. For example, mobile assistance systems could use SDA-M-based analyses to suggest training goals, provide athlete- and sport-specific feedback, and track practitioners' learning curves in terms of how task-related memory structures develop over time.
- Finally, one of the most important questions for industrial applications would be how algorithmic SDA-M can be applied to more complex activities that comprise much more than 15-20 basic action steps. A possible approach might be to divide such activities hierarchically into sub-tasks and create separate split procedures to isolate the respective basic actions for each of these sub-tasks. Additionally, a distinct split procedure could assess how the overarching activities are mentally represented and structured in terms of their sub-tasks. This approach would vastly reduce the number of comparisons and therefore the total time required for split procedures. Future research could assess the practical suitability of this approach, or discard it and find a better one.

  After all, as Don Norman noted at the 1996 Annual Conference of the Travel and Tourism Research Association (U.S.):

  "Academics get paid for being clever, not for being right."

Appendix

The following sections serve as a dump for additional information that did not fit into the main body of the thesis but might be of interest to some readers.

A Persona creation procedure outline

```
Replace NAs in raw data by column medians

Calculate PCA of the standardized data

Scree test,
retain first 3 principal components (PCs)

Varimax rotation

Invert loadings iff this results in a higher sum
of the absolute values of all loadings
/* rationale: more interest in knowing what
stakeholder groups need than in what they don't
need */

Calculate factor scores for each participant

Identify N<4 best representatives for each PC,
i.e. respondents with highest scores for that PC
and low(er) scores for all other PCs

Calculate mean values for these representative
subgroups and compare them to the respective PC's
factor loadings
```

```
  /* rationale: put absolute values of high-loading
     variables in relation to other variables and
     identify additional important (non-idiosyncratic)
     features */

Base persona descriptions on:
- specific loadings
- mean values of the subgroup

Double-checks:
- keep different number of PCs
- include/exclude demographic data in addition to
  feature ratings
```

B Action descriptions for user test scenario “coffee”

Since all participants in this user test scenario were German native speakers, action descriptions in German have been used for SDA-M, which translate to English as follows:

```
a1: Reading display instruction "Fill water"
a2: Unscrewing the cistern lid
a3: Opening the door
a4: Pulling out the container at the bottom left
a5: Pulling out the container at the bottom right
a6: Pulling out the container at the top right
a7: Filling in water and inserting the container
a8: Closing the door
a9: Placing the coffee cup under the central spout
a10: Selecting drink
```

The correct sequence of action IDs for the test task was:
$(a_1, a_3, a_5, a_7, a_8, a_9, a_{10})$.

C Further contributions

Publications

- Neumann, A., **Strenge**, B., Uhlich, J., Schlicher, K., Maier, G. W., Schalkwijk, L., Waßmuth, J., et al. (2020). AVIKOM: Towards a mobile audiovisual cognitive assistance system for modern manufacturing and logistics. *PETRA '20: Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments* New York: ACM. doi:10.1145/3389189.3389191

- Blattgerste, J., Renner, P., **Strenge**, B., & Pfeiffer, T. (2018). In-Situ Instructions Exceed Side-by-Side Instructions in Augmented Reality Assisted Assembly. *PETRA '18: Proceedings of the 11th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 133-140. New York, NY, USA: ACM. doi:10.1145/3197768.3197778

- Essig, K., **Strenge**, B., & Schack, T. (2018). Assistierende Technologie zur Förderung beruflichen Entwicklungspotenzials. In G. W. Maier, G. Engels, & E. Steffen (Eds.), Springer Reference Psychologie. *Handbuch Gestaltung digitaler und vernetzter Arbeitswelten* (Living reference work, continuously updated edition, pp. 1-29). Berlin, Heidelberg: Springer. doi:10.1007/978-3-662-52903-4_21-1

- Essig, K., **Strenge**, B., & Schack, T. (2018). Die intelligente ADAMAAS-Datenbrille – Chancen und Risiken des Einsatzes mobiler Assistiver Technologien für die Inklusion. In A. Burchardt & H. Uszkoreit (Eds.), *IT für soziale Inklusion. Digitalisierung – Künstliche Intelligenz – Zukunft für alle* (pp. 33-40). Berlin, Boston: De Gruyter. doi:10.1515/9783110561371-004

- **Strenge**, B., Vogel, L., & Schack, T. (2018). Individualized cognitive assistance by smart glasses for manual assembly processes in industry. In R. Weidner (Ed.), *Proceedings of the 3rd Transdisciplinary Conference on Support Technologies (TCST18)* (pp. 399-407). Hamburg: Helmut-Schmidt-Universität.

- Essig, K., **Strenge**, B., Frank, C., & Schack, T. (2018). Neue Untersuchungs- und Trainingsmethoden für den Sportbereich durch den Einsatz von modernen multi-modalen Blickerfassungs- und Feedbacksystemen. In U. Borges, L. Bröker, S. Hoffmann, T. Hosang, S. Laborde, R. Liepelt, B. Lobinger, et al. (Eds.), *Die Psychophysiologie der Handlung*. 50. Jahrestagung der Arbeitsgemeinschaft für Sportpsychologie (asp) in Köln (p. 50).

- Blattgerste, J., **Strenge**, B., Renner, P., Pfeiffer, T., & Essig, K. (2017). Comparing Conventional and Augmented Reality Instructions for Manual Assembly Tasks. *PETRA'17: Proceedings of the 10th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 75-82. New York, NY, USA: ACM. doi:10.1145/3056540.3056547

- Wittmaack, L., Esslinger, B., Schmidt, L., **Strenge**, B., & Wacker, A. (2016). Vertrauensvolle E-Mail-Kommunikation. *Datenschutz und Datensicherheit - DuD*, 40(5), pp. 271–277. doi:10.1007/s11623-016-0595-9

- Esslinger, B., Schmidt, L., **Strenge**, B., & Wacker, A. (2016). Unpopuläre E-Mail-Verschlüsselung – Auflösung des Henne-Ei-Problems. *Datenschutz und Datensicherheit - DuD*, 40(5), pp. 283–289. doi:10.1007/s11623-016-0597-7

- Essig, K., **Strenge**, B., & Schack, T. (2016). ADAMAAS – Towards Smart Glasses for Mobile and Personalized Action Assistance. *PETRA'16: Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 46:1-46:4. New York, NY, USA: ACM. doi:10.1145/2910674.2910727

- **Strenge**, B., Sieburg, S., & Schmidt, L. (2016). Experimental Comparison of Sidestick Steering Configurations for an Innovative Electric Two-wheel Vehicle. In B. Deml, P. Stock, R. Bruder, & C. M. Schlick (Eds.), *Advances in Ergonomic Design of Systems, Products and Processes*: Proceedings of the Annual Meeting of GfA 2015 (pp. 313–326). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-662-48661-0_21

- Fischer, H., **Strenge**, B., & Nebe, K. (2013). Towards a Holistic Tool for the Selection and Validation of Usability Method Sets Supporting Human-Centered Design. In A. Marcus (Ed.), *Lecture Notes in Computer Science: Vol. 8012. Design, User Experience, and Usability. Design Philosophy, Methods, and Tools* (pp. 252-261). Berlin, Heidelberg: Springer Science + Business Media. doi:10.1007/978-3-642-39229-0_28

Teaching

- Supervision of five Master's theses in the "Intelligence and Motion" study program (two of them in industry cooperation with automotive suppliers)

- Supervision of two Bachelor's theses in the "Cognitive Informatics" study program

- Course *"User Experience Engineering and Assistive Systems"* (2018)

- Course *"Einführung Programmiersprachen (MATLAB I)"* (2017 – 2019)

- Course *"Einführung Datenanalyse (MATLAB II)"* (2017 – 2019)

- Course *"Kognitive Systeme: Mobile and Stationary Assistive Systems and User Experience Engineering"* (2016 – 2017)

List of Acronyms

ACT-R	Adaptive Control of Thought – Rational
ADAMAAS	Adaptive and Mobile Action Assistance
AMPA	Analysis of Most Probable Actions
AR	Augmented Reality
ARI	Adjusted Rand Index
AWOSE	Agile Worth-Oriented Systems Engineering
BAC	Basic Action Concept
CAA-A	Cognitive Action Architecture Approach
CASPA	Correct Action Selection Probability Analysis
CASPA$_d$	CASPA with default threshold
CASPA$_i$	CASPA with informed threshold
CASPA$_m$	CASPA mean
CI	Correctly Instructed
CIT	Cognitive Interaction Technology
CITEC	Center of Excellence in Cognitive Interaction Technology
CMPA	Correct Most-Probable Action
E-SDLC	Ethical System Design Lifecycle
EI	Erroneously Instructed

ELSI	Ethical, Legal and Social Implications
ICT	Information and Communication Technology
MEESTAR	Model for Ethical Evaluation of Socio-Technical Arrangements
NPV	Negative Predictive Value
PC	Personal Computer *or* Principal Component
PCA	Principal Components Analysis
PPV	Positive Predictive Value
SD	Standard Deviation
SDA-M	Structural-Dimensional Analysis of Mental Representations
SMC	Simple Matching Coefficient
SUS	System Usability Scale
UCD	User-Centred Design
UE	Usability Engineering
UI	User Interface
UML	Unified Modeling Language
UX	User Experience
VR	Virtual Reality
VSD	Value-Sensitive Design
WCD	Worth-Centered Development
XP	Extreme Programming

List of Tables

List of Figures

Bibliography

Alkhatib, O. J., & Abdou, A. (2017). An ethical (descriptive) framework for judgment of actions and decisions in the construction industry and engineering–part i. *Science and Engineering Ethics*. doi: 10.1007/s11948-017-9895-1

Alphabet Inc. (2017). *Code of conduct.* Retrieved 2020-05-11, from `https://abc.xyz/investor/other/code-of-conduct/`

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological review*, *89*(4), 369.

Anderson, J. R. (1983). *The architecture of cognition*. Psychology Press.

Anderson, J. R. (1993). *Rules of the mind*. Psychology Press.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press. Retrieved from `https://books.google.de/books?id=7jpnDAAAQBAJ`

Anderson, J. R. (2020). *Cognitive psychology and its implications* (9th ed.). New York, NY: Worth Publishers.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, *8*(4), 629–647. doi: 10.3758/BF03196200

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Anderson, J. R., Fincham, J. M., Qin, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences*, *12*(4), 136 - 143. doi: 10.1016/j.tics.2008.01.006

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, N.J.: Lawrence Erlbaum Associates. Paperback.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

Bairaktarova, D., & Woodcock, A. (2017). Engineering student's ethical awareness and behavior: A new motivational model. *Science and Engineering Ethics*, *23*(4), 1129–1157. doi: 10.1007/s11948-016-9814-x

Bangor, A., Kortum, P., & Miller, J. (2009, May). Determining what individual sus scores mean: Adding an adjective rating scale. *J. Usability Studies*, *4*(3), 114–123.

Batagelj, V., & Bren, M. (1995). Comparing resemblance measures. *Journal of classification*, *12*, 73–90. doi: 10.1007/BF01202268

Beauchamp, T. L., & Childress, J. F. (2006). *Principles of biomedical ethics*. Oxford University Press.

Beauregard, R., & Corriveau, P. (2007). User experience quality: A conceptual framework for goal setting and measurement. In V. G. Duffy (Ed.), *Digital human modeling* (pp. 325–332). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-73321-8_38

Beck, K. (2000). *Extreme programming explained: embrace change*. Addison-Wesley.

Behrmann, E., & Rauwald, C. (2016). *Mercedes boots robots from the production line.* Retrieved 25 Nov 2019, from `https://www.bloomberg.com/news/articles/2016-02-25/why-mercedes-is-halting-robots-reign-on-the-production-line`

Belavkin, R. V. (2001). The role of emotion in problem solving. In *Proceedings of the AISB'01 symposium on emotion, cognition and affective computing* (pp. 49–57). Heslington, York, England.

Bernstein, N. A. (1967). *The co-ordination and regulation of movements*. Pergamon Press Ltd.

Bevan, N. (2003). Usabilitynet methods for user centred design. In *Human-computer interaction: theory and practice* (Vol. 1, pp. 434–438). Lawrence Erlbaum.

Blattgerste, J., Renner, P., & Pfeiffer, T. (2018). Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views. In *Proceedings of the workshop on communication by gaze interaction.* New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3206343.3206349

Blattgerste, J., Renner, P., Strenge, B., & Pfeiffer, T. (2018). In-situ instructions exceed side-by-side instructions in augmented reality assisted assembly. In *Proceedings of the 11th pervasive technologies related to assistive environments conference (PETRA '18)* (pp. 133–140). New York, NY, USA: ACM. doi: 10.1145/3197768.3197778

Blattgerste, J., Strenge, B., Renner, P., Pfeiffer, T., & Essig, K. (2017). Comparing conventional and augmented reality instructions for manual assembly tasks. In *Proceedings of the 10th international conference on pervasive technologies related to assistive environments (PETRA '17)* (pp. 75–82). New York, NY, USA: ACM. doi: 10.1145/3056540.3056547

Bläsing, B., Tenenbaum, G., & Schack, T. (2009). The cognitive structure of movements in classical dance. *Psychology of Sport and Exercise*, *10*(3), 350–360. doi: 10.1016/j.psychsport.2008.10.001

Boenink, M., Swierstra, T., & Stemerding, D. (2010). Anticipating the interaction between technology and morality: A scenario study of experimenting with humans in bionanotechnology. *Studies in Ethics, Law, and Technology*, *4*(2). doi: 10.2202/1941-6008.1098

Borenstein, J., Herkert, J. R., & Miller, K. W. (2017). Self-driving cars and engineering ethics: The need for a system level analysis. *Science and Engineering Ethics*, *25*(2), 383–398. doi: 10.1007/s11948-017-0006-0

Braun, S. M., Beurskens, A. J., Schack, T., Marcellis, R. G., Oti, K. C., Schols, J. M., & Wade, D. T. (2007). Is it possible to use the structural dimension analysis of motor memory (SDA-M) to investigate representations of motor actions in stroke patients? *Clinical Rehabilitation*, *21*(9), 822–832.

Brickey, J., Walczak, S., & Burgess, T. (2012, May). Comparing semi-automated clustering methods for persona development. *IEEE Transactions on Software Engineering*, *38*(3), 537-546. doi: 10.1109/TSE.2011.60

Brief, A. P., Aldag, R. J., & Chacko, T. I. (1977, dec). The miner sentence completion scale: An appraisal. *Academy of Management Journal*, *20*(4), 635–643. doi: 10.5465/255362

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, Aug). The balanced accuracy and its posterior distribution. In *20th international conference on pattern recognition (icpr), 2010* (p. 3121-3124). doi: 10.1109/ICPR.2010.764

Brooke, J. (1996). Sus: A 'quick and dirty' usability scale. In *Usability evaluation in industry* (p. 189-194). Taylor and Francis.

Buchanan, B. G., Davis, R., & Feigenbaum, E. A. (2018). Expert systems: A perspective from computer science. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The cambridge handbook of expertise and expert performance* (2nd ed., pp. 84–104). Cambridge University Press.

Butler, K. A. (1985). Connecting theory and practice: A case study of achieving usability goals. *ACM SIGCHI Bulletin*, *16*(4), 85–88. doi: 10.1145/1165385.317472

Büttner, S., Mucha, H., Funk, M., Kosch, T., Aehnelt, M., Robert, S., & Röcker, C. (2017). The design space of augmented and virtual reality applications for assistive environments in manufacturing: A visual approach. In *Proceedings of the 10th international conference on pervasive technologies related to assistive environments (PETRA '17)* (pp. 433–440). New York, NY, USA: ACM. doi: 10.1145/3056540.3076193

Carpenter, W. B. (1852). *On the influence of suggestion in modifying and directing muscular movement, independently of volition.*

Chapman, C. N., & Milham, R. P. (2006). The personas' new clothes: Methodological and practical arguments against a popular method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(5), 634-636. doi: 10.1177/154193120605000503

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Cheruvalath, R. (2017). Does studying 'ethics' improve engineering students' meta-moral cognitive skills? *Science and Engineering Ethics*, *25*(2), 583–596. doi: 10.1007/s11948-017-0009-x

Cockton, G. (2005). A development framework for value-centred design. In *Chi '05 extended abstracts on human factors in computing systems* (pp. 1292–1295). New York: ACM. doi: 10.1145/1056808.1056899

Cockton, G. (2006). Designing worth is worth designing. In *Proceedings of the 4th nordic conference on human-computer interaction changing roles - NordiCHI '06.* ACM Press. doi: 10.1145/1182475.1182493

Cockton, G. (2008a). Feature: Designing worth—connecting preferred means to desired ends. *Interactions*, *15*(4), 54–57. doi: 10.1145/1374489.1374502

Cockton, G. (2008b). Putting value into e-valu-ation. In E. L. C. Law, E. T. Hvannberg, & G. Cockton (Eds.), *Maturing usability* (pp. 287–317). London: Springer. doi: 10.1007/978-1-84628-941-5_13

Cockton, G. (2008c). What worth measuring is. In *Proceedings of the international workshop on meaningful measures: Valid useful user experience measurement (VUUM 2008)* (pp. 60–66). Institute of Research in Informatics of Toulouse (IRIT).

Cockton, G. (2009a). Getting there: Six meta-principles and interaction design. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2223–2232). New York: ACM. doi: 10.1145/1518701.1519041

Cockton, G. (2009b). When and why feelings and impressions matter in interaction design. In *Proceedings of the conference: Interfejs użytkownika - kansei w praktyce* (pp. 7–31). Wydawnictwo PJWSTK.

Cockton, G. (2012). Making designing worth worth designing. In *ACM SIGCHI conference on human factors in computing systems (CHI'12)*. ACM. Retrieved from `http://nrl.northumbria.ac.uk/11838/1/cockton.pdf`

Cockton, G., Kirk, D., Sellen, A., & Banks, R. (2009). Evolving and augmenting worth mapping for family archives. In *Proceedings of the 23rd british hci group annual conference on people and computers: Celebrating people and technology* (pp. 329–338). London: British Computer Society. Retrieved from `http://dl.acm.org/citation.cfm?id=1671011.1671053`

Cockton, G., Kujala, S., Nurkka, P., & Hölttä, T. (2009). Supporting worth mapping with sentence completion. In *Human-computer interaction – INTERACT 2009* (pp. 566–581). Springer. doi: 10.1007/978-3-642-03658-3_61

Cooper, A. (2004). *The inmates are running the asylum*. Indianapolis: SAMS.

d'Avella, A., Giese, M., Ivanenko, Y. P., Schack, T., & Flash, T. (2015). Editorial: Modularity in motor control: from muscle synergies to cognitive action representation. *Frontiers in Computational Neuroscience*, *9*, 126. doi: 10.3389/fncom.2015.00126

Davis, J., & Nathan, L. P. (2015). Value sensitive design: Applications, adaptations, and critiques. In *Handbook of ethics, values, and technological design* (pp. 11–40). Springer Netherlands. doi: 10.1007/978-94-007-6970-0_3

Donoghue, S. (2000). Projective techniques in consumer research. *Journal of Family Ecology and Consumer Sciences*, *28*(1), 47–53. doi: 10.4314/jfecs.v28i1.52784

Ericsson, K. A. (2007). Deliberate practice and the modifiability of body and mind: Toward a science of the structure and acquisition of expert and elite performance. *International Journal of Sport Psychology*, *38*(1), 4–34.

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine*, *15*(11), 988-994. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1553-2712.2008.00227.x` doi: 10.1111/j.1553-2712.2008.00227.x

Essig, K., Strenge, B., & Schack, T. (2016). ADAMAAS: Towards smart glasses for mobile and personalized action assistance. In *Proceedings of the 9th ACM international conference on pervasive technologies related to assistive environments (PETRA '16)* (pp. 46:1–46:4). New York, NY, USA: ACM. doi: 10.1145/2910674.2910727

Evans, G., Miller, J., Pena, M. I., MacAllister, A., & Winer, E. (2017). Evaluating the microsoft hololens through an augmented reality assembly application. In *Degraded environments: Sensing, processing, and display 2017* (Vol. 10197, pp. 101970V-1–101970V-16). Retrieved 26 Nov 2019, from `https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1178&context=me_conf` doi: 10.1117/12.2262626

Fischer, H., Strenge, B., & Nebe, K. (2013). Towards a holistic tool for the selection and validation of usability method sets supporting human-centered design. In A. Marcus (Ed.), *Design, user experience, and usability. design philosophy, methods, and tools* (Vol. 8012, pp. 252–261). Berlin: Springer.

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.

Frank, C., Land, W. M., Popp, C., & Schack, T. (2014, 04). Mental representation and mental practice: Experimental investigation on the functional links between motor memory and motor imagery. *PLOS ONE*, *9*(4), 1-12. doi: 10.1371/journal.pone.0095175

Frank, C., Land, W. M., & Schack, T. (2013). Mental representation and learning: The influence of practice on the development of mental representation structure in complex action. *Psychology of Sport and Exercise*, *14*(3), 353–361.

Frank, C., Land, W. M., & Schack, T. (2016). Perceptual-cognitive changes during motor learning: The influence of mental and physical practice on mental representation, gaze behavior, and performance of a complex action. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01981

Friedman, B., Hendry, D. G., & Borning, A. (2017). A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, *11*(2), 63–125. doi: 10.1561/1100000015

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value sensitive design and information systems. In *The handbook of information and computer ethics* (pp. 69–101). John Wiley & Sons, Inc. doi: 10.1002/9780470281819.ch4

Funakoshi, G. (1938). Karate-do nijukkajo to sono kaishaku. In *Karate-do taikan.*

Funakoshi, G., Nakasone, G., & Takagi, J. (2003). *The twenty guiding principles of karate*. Kodansha International.

Funk, M., Bächler, A., Bächler, L., Kosch, T., Heidenreich, T., & Schmidt, A. (2017). Working with augmented reality?: A long-term analysis of in-situ instructions at the assembly workplace. In *Proceedings of the 10th international conference on pervasive technologies related to assistive environments* (pp. 222–229). New York, NY, USA: ACM. doi: 10.1145/3056540.3056548

Funk, M., Kosch, T., Greenwald, S. W., & Schmidt, A. (2015). A benchmark for interactive augmented reality instructions for assembly tasks. In *Proceedings of the 14th international conference on mobile and ubiquitous multimedia* (pp. 253–257). New York, NY, USA: ACM. doi: 10.1145/2836041.2836067

Furniss, D., Blandford, A., & Curzon, P. (2007). Usability evaluation methods in practice: Understanding the context in which they are embedded. In *Proceedings of the 14th european conference on cognitive ergonomics: Invent! explore!* (p. 253–256). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1362550.1362602

Gelfand, S. D. (2016). Using insights from applied moral psychology to promote ethical behavior among engineering students and professional engineers. *Science and Engineering Ethics*, *22*(5), 1513–1534. doi: 10.1007/s11948-015-9721-6

Hagan Jr., J. E., Schack, T., & Koester, D. (2018). Passion play. embracing new scientific perspectives for improved sport psychology consulting. *SOJ Psychology*, *4*(3), 1–5.

Harbour, R., & Scemama, S. (2017). *Surprise: Robots aren't replacing humans in key areas of manufacturing.* Retrieved 25 Nov 2019, from `https://www.forbes.com/sites/oliverwyman/2017/02/03/surprise-the-correct-answer-is-not-always-to-go-with-the-robot-just-ask-some-automakers/`

Heinen, T., & Schack, T. (2004). Bewegungsgedächtnis und Bewegungsausführung–Optimierung von Rotationsbewegungen im Gerätturnen. *Lehren und Lernen im Turnen. Veröffentlichungsband zur Jahrestagung*, 85–95.

Heinen, T., & Schwaiger, J. (2002). Optimierung des Trainingsprozesses im Kunstturnen durch kognitive Verfahren. *Expertise im Sport: Lehren, lernen, leisten*, 67–68.

Heinen, T., Schwaiger, J., & Schack, T. (2002). Optimising gymnastics training with cognitive methods. In *Proceedings of 7th annual congress of the european college of sport science* (p. 608).

Herzberg, F. (1964). The motivation-hygiene concept and problems of manpower. *Personnel Administration*, *27*, 3–7.

Hinton, G. E. (2007). Boltzmann machine. *Scholarpedia*, *2*(5), 1668. (revision 91075)

Hodges, N. J., Huys, R., & Starkes, J. L. (2007). Methodological review and evaluation of research in expert performance in sport. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (pp. 161–183). John Wiley & Sons.

Hoffmann, J. (2003). Anticipatory behavioral control. In *Anticipatory behavior in adaptive learning systems* (pp. 44–65). Springer.

Hofmann, B. (2017). Toward a method for exposing and elucidating ethical issues with human cognitive enhancement technologies. *Science and Engineering Ethics*, *23*(2), 413–429. doi: 10.1007/s11948-016-9791-0

Hofmann, B., Haustein, D., & Landeweerd, L. (2017). Smart-glasses: Exposing and elucidating the ethical issues. *Science and Engineering Ethics*, *23*(3), 701–721. doi: 10.1007/s11948-016-9792-z

Holaday, M., Smith, D. A., & Sherry, A. (2000). Sentence completion tests: A review of the literature and results of a survey of members of the society for personality assessment. *Journal of Personality Assessment*, *74*(3), 371–383. doi: 10.1207/s15327752jpa7403_3

Holzinger, A., Errath, M., Searle, G., Thurnher, B., & Slany, W. (2005). From extreme programming and usability engineering to extreme usability in software engineering education. In *29th annual international computer software and applications conference (COMPSAC'05)*. IEEE. doi: 10.1109/compsac.2005.80

Hossner, E.-J., Schiebl, F., & Göhner, U. (2015). A functional approach to movement analysis and error identification in sports and physical education. *Frontiers in Psychology*, *6*, 1339. doi: 10.3389/fpsyg.2015.01339

Hu, M., Akella, P., Kapoor, B., & Prager, D. (2018). *The state of human factory analytics*. Retrieved 25 Nov 2019, from `https://drishti.com/state-of-human-factory-analytics/`

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218. doi: 10.1007/BF01908075

Hülsmann, F., Frank, C., Senna, I., Ernst, M. O., Schack, T., & Botsch, M. (2019). Superimposed skilled performance in a virtual mirror improves motor performance and cognitive representation of a full body motor action. *Frontiers in Robotics and AI*, *6*, 43. doi: 10.3389/frobt.2019.00043

Ienca, M., Wangmo, T., Jotterand, F., Kressig, R. W., & Elger, B. (2017). Ethical design of intelligent assistive technologies for dementia: A descriptive review. *Science and Engineering Ethics*, *24*(4), 1035–1055. doi: 10.1007/s11948-017-9976-1

ISO 14915. Software ergonomics for multimedia user interfaces (1st ed.) (Norm No. ISO 14915). (2002).

ISO 15005. Road vehicles — Ergonomic aspects of transportation and control systems — Dialogue management principles and compliance procedures (2nd ed.) (Norm No. ISO 15005:2017). (2017).

ISO 9241. Ergonomics of human-system interaction — Part 110: Dialogue principles (1st ed.) (Norm No. ISO 9241-110:2006). (2006).

ISO 9241. Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts (2nd ed.) (Norm No. ISO 9241-11:2018). (2018).

ISO 9241. Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems (2nd ed.) (Norm No. ISO 9241-210:2019). (2019).

ISO/TR 16982. Ergonomics of human-system interaction — Usability methods supporting human-centred design (1st ed.) (Norm No. ISO/TR 16982:2002). (2002).

Jacksteit, R., Mau-Moeller, A., Behrens, M., Bader, R., Mittelmeier, W., Skripitz, R., & Stöckel, T. (2017). The mental representation of the human gait in patients with severe knee osteoarthrosis: a clinical study to aid understanding of impairment and disability. *Clinical Rehabilitation*. doi: 10.1177/0269215517719312

James, W. (1890). *The principles of psychology*. Henry Holt and Company.

Jeannerod, M. (2004). Actions from within. *International Journal of Sport and Exercise Psychology*, *2*(4), 376–402.

Jeraj, D., Musculus, L., & Lobinger, B. (2017, December/2017). Body image and mental representation in table tennis players who do versus do not use a prosthesis. *Problems of Psychology in the 21st Century*, *11*, 22–30. Retrieved from `http://oaji.net/articles/2017/444-1515690478.pdf`

Kaeser, J. (2017). *Why robots will improve manufacturing jobs.* Retrieved 25 Nov 2019, from `https://time.com/4940374/joe-kaeser-siemens-robots-jobs/`

Kermisch, C., & Depaus, C. (2018). The strength of ethical matrixes as a tool for normative analysis related to technological choices: The case of geological disposal for radioactive waste. *Science and Engineering Ethics*, *24*(1), 29–48. doi: 10.1007/s11948-017-9882-6

Kim, T., Frank, C., & Schack, T. (2017). A systematic investigation of the effect of action observation training and motor imagery training on the development of mental representation structure and skill performance. *Frontiers in Human Neuroscience*, *11*, 499. doi: 10.3389/fnhum.2017.00499

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*, 17–94. doi: 10.1007/s10462-018-9646-y

Laird, J. E. (2012). *The soar cognitive architecture*. MIT press.

Land, W., Frank, C., & Schack, T. (2014). The influence of attentional focus on the development of skill representation in a complex action. *Psychology of Sport and Exercise*, *15*(1), 30–38. doi: 10.1016/j.psychsport.2013.09.006

Land, W., Volchenkov, D., Bläsing, B., & Schack, T. (2013). From action representation to action execution: exploring the links between cognitive and biomechanical levels of motor control. *Frontiers in Computational Neuroscience*, *7*, 127. doi: 10.3389/fncom.2013.00127

Lander, H.-J. (1991). Ein methodischer Ansatz zur Ermittlung der Struktur und der Dimensionierung einer intern repräsentierten Begriffsmenge. *Zeitschrift für Psychologie*, *199*(2), 167–176.

Lander, H.-J., & Lange, K. (1992). Eine differentialpsychologische Analyse begrifflich-strukturierten Wissens. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*, *200*(3), 181–197.

Lebière, C., Anderson, J. R., & Reder, L. M. (1994). Error modeling in the ACT-R production system. In *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 555–559).

Lee, J. C., McCrickard, D. S., & Stevens, K. T. (2009). Examining the foundations of agile usability with eXtreme scenario-based design. In *2009 agile conference.* IEEE. doi: 10.1109/agile.2009.30

Leese, M. (2017). Holding the project accountable: Research governance, ethics, and democracy. *Science and Engineering Ethics*, *23*(6), 1597–1616. doi: 10.1007/s11948-016-9866-y

Lewin, K. (1945). The research center for group dynamics at massachusetts institute of technology. *Sociometry*, *8*(2), 126–136. Retrieved from http://www.jstor.org/stable/2785233

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In *Proceedings of the 1st international conference on human centered design: Held as part of hci international 2009* (p. 94–103). Berlin, Heidelberg: Springer-Verlag.

Lex, H., Essig, K., Knoblauch, A., & Schack, T. (2015). Cognitive representations and cognitive processing of team-specific tactics in soccer. *PloS one*, *10*(2), e0118219.

Lokhorst, G.-J. C. (2018). Martin peterson: The ethics of technology: A geometric analysis of five moral principles. *Science and Engineering Ethics*, *24*(5), 1641–1643. doi: 10.1007/s11948-017-0014-0

Lotze, H. (1852). *Medicinische Psychologie oder Psychologie der Seele*. Weidmann.

Louie, A. (2010). Robert rosen's anticipatory systems. *Foresight*, *12*(3), 18–29.

Macnamara, B. N., Moreau, D., & Hambrick, D. Z. (2016). The relationship between deliberate practice and performance in sports: A meta-analysis. *Perspectives on Psychological Science*, *11*(3), 333-350. doi: 10.1177/1745691616635591

Maguire, E. A., Woollett, K., & Spiers, H. J. (2006). London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus*, *16*(12), 1091-1101. doi: 10.1002/hipo.20233

Manders-Huits, N. (2011). What values in design? the challenge of incorporating moral values into design. *Science and Engineering Ethics*, *17*(2), 271–287. doi: 10.1007/s11948-010-9198-2

Manzeschke, A. (2015). MEESTAR–ein Modell angewandter Ethik im Bereich assistiver Technologien. In K. Weber, D. Frommeld, A. Manzeschke, & H. Fangerau (Eds.), *Technisierung des Alters–Beitrag zu einem guten Leben* (pp. 263–283). Stuttgart: Franz Steiner Verlag.

Manzeschke, A., Weber, K., Rother, E., & Fangerau, H. (2015). *Ethical questions in the area of age appropriate assisting systems*. German Federal Ministry of Education and Research, VDI/VDE Innovation+ Technik GmbH.

Mao, J.-Y., Vredenburg, K., Smith, P. W., & Carey, T. (2001). User-centered design methods in practice: A survey of the state of the art. In *Proceedings of the 2001 conference of the centre for advanced studies on collaborative research* (p. 12). IBM Press.

Maycock, J., Dornbusch, D., Elbrechter, C., Haschke, R., Schack, T., & Ritter, H. (2010). Approaching manual intelligence. *KI-Künstliche Intelligenz*, *24*(4), 287–294.

Mayhew, D. J. (1999). *The usability engineering lifecycle*. Burlington: Morgan Kaufmann.

McGinn, J., & Kotamraju, N. (2008). Data-driven persona development. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1521–1524). New York: ACM. doi: 10.1145/1357054.1357292

McKinlay, R. (2016). Technology: Use or lose our navigation skills. *Nature*, *531*(7596), 573–575.

Meier, C., Frank, C., Gröben, B., & Schack, T. (2020). Verbal instructions and motor learning: How analogy and explicit instructions influence the development of mental representations and tennis serve performance. *Frontiers in Psychology*, *11*, 2. doi: 10.3389/fpsyg.2020.00002

Memmel, T., Gundelsweiler, F., & Reiterer, H. (2007). Agile human-centered software engineering. BCS Learning & Development. doi: 10.14236/ewic/hci2007.17

Miaskiewicz, T., Sumner, T., & Kozar, K. A. (2008). A latent semantic analysis methodology for the identification and creation of personas. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1501–1510). New York: ACM. doi: 10.1145/1357054.1357290

Miñano, R., Uruburu, Á., Moreno-Romero, A., & Pérez-López, D. (2016). Strategies for teaching professional ethics to IT engineering degree students and evaluating the result. *Science and Engineering Ethics*, *23*(1), 263–286. doi: 10.1007/s11948-015-9746-x

Mura, M. D., Dini, G., & Failli, F. (2016). An integrated environment based on augmented reality and sensing device for manual assembly workstations. *Procedia CIRP*, *41*, 340 - 345. (Research and Innovation in Manufacturing: Key Enabling Technologies for the Factories of the Future - Proceedings of the 48th CIRP Conference on Manufacturing Systems) doi: 10.1016/j.procir.2015.12.128

Murphy, C., & Gardoni, P. (2017). Understanding engineers' responsibilities: A prerequisite to designing engineering education. *Science and Engineering Ethics*. doi: 10.1007/s11948-017-9949-4

Musk, E. (2018). *Elon Musk on Twitter: "Yes, excessive automation at Tesla was a mistake. To be precise, my mistake. Humans are underrated.".* Retrieved 25 Nov 2019, from `https://twitter.com/elonmusk/status/984882630947753984`

Nielsen, J. (1993). *Usability engineering*. San Francisco and CA and USA: Morgan Kaufmann Publishers Inc.

Nielsen, J. (2005). *10 usability heuristics for user interface design.* Retrieved February 22, 2017, from `https://www.nngroup.com/articles/ten-usability-heuristics/`

Nitsch, J. R. (2004). Die handlungstheoretische Perspektive: ein Rahmenkonzept für die sportpsychologische Forschung und Intervention. *Zeitschrift für Sportpsychologie*, *11*(1), 10–23.

Norman, D. A. (1981). Categorization of action slips. *Psychological review*, *88*(1), 1.

Norman, D. A. (2004). *Ad-hoc personas & empathetic focus.* Retrieved 12 Sep 2019, from `https://jnd.org/ad-hoc_personas_empathetic_focus`

Norman, D. A. (2014). *Things that make us smart: Defending human attributes in the age of the machine*. Diversion Books.

Obendorf, H., & Finck, M. (2008). Scenario-based usability engineering techniques in agile development processes. In *Proceeding of the twenty-sixth annual CHI conference extended abstracts on human factors in computing systems - CHI'08.* ACM Press. doi: 10.1145/1358628.1358649

Olson, J. S., & Moran, T. P. (1995). Mapping the method muddle: Guidance in using methods for user interface design. In *Proceedings of a workshop on human-computer interface design: Success stories, emerging methods, and real-world context: Success stories, emerging methods, and real-world context* (p. 269–300). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Polson, P. G., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, *36*(5), 741–773. doi: 10.1016/0020-7373(92)90039-N

Prinz, W. (1997). Perception and action planning. *European journal of cognitive psychology*, *9*(2), 129–154.

Pruitt, J., & Adlin, T. (2006). *The persona lifecycle: Keeping people in mind throughout product design*. Amsterdam: Elsevier.

Pruitt, J., & Grudin, J. (2003). Personas: Practice and theory. In *Proceedings of the 2003 conference on designing for user experiences - DUX'03*. ACM. doi: 10.1145/997078.997089

Quesenbery, W. (2001). What does usability mean: Looking beyond 'ease of use'. In *Proceedings of the 48th annual conference, society for technical communication.* Retrieved from `http://www.wqusability.com/articles/more-than-ease-of-use.html`

Quesenbery, W. (2003). The five dimensions of usability. In M. J. Albers & M. B. Mazur (Eds.), *Content and complexity: Information design in technical communication* (pp. 81–102). Routledge.

Reason, J. (1990). *Human error*. Cambridge university press.

Reijers, W., Wright, D., Brey, P., Weber, K., Rodrigues, R., O'Sullivan, D., & Gordijn, B. (2017). Methods for practising ethics in research and innovation: A literature review, critical analysis and recommendations. *Science and Engineering Ethics*, *24*(5), 1437–1481. doi: 10.1007/s11948-017-9961-8

Renner, P., Blattgerste, J., & Pfeiffer, T. (2018). A path-based attention guiding technique for assembly environments with target occlusions. In *IEEE virtual reality 2018*. IEEE. Retrieved from `https://pub.uni-bielefeld.de/record/2917385`

Renner, P., & Pfeiffer, T. (2017). Attention guiding techniques using peripheral vision and eye tracking for feedback in augmented-reality-based assistance systems. In *2017 IEEE symposium on 3D user interfaces (3DUI)* (p. 186-194). doi: 10.1109/3DUI.2017.7893338

Renner, P., & Pfeiffer, T. (2017a). Augmented reality assistance in the central field-of-view outperforms peripheral displays for order picking: Results from a virtual reality simulation study. In *ISMAR 2017*. IEEE. doi: 10.1109/ISMAR-Adjunct.2017.59

Renner, P., & Pfeiffer, T. (2017b). Evaluation of attention guiding techniques for augmented reality-based assistance in picking and assembly tasks. In *Proceedings of the 22nd international conference on intelligent user interfaces companion* (pp. 89–92). New York, NY, USA: ACM. doi: 10.1145/3030024.3040987

Renner, P., & Pfeiffer, T. (2017c). Eye-tracking-based attention guidance in mobile augmented reality assistance systems. In R. Radach, H. Deubel, C. Vorstius, & M. J. Hofmann (Eds.), *Abstracts of the 19th european conference on eye movements* (Vol. 10, p. 218). Retrieved from https://pub.uni-bielefeld.de/record/2917484 doi: 10.16910/jemr.10.6

Ritter, H. (2010). Cognitive interaction technology. *KI - Künstliche Intelligenz*, *24*(4), 319–322. doi: 10.1007/s13218-010-0063-x

Ritter, H., & Sagerer, G. (2009). Excellence cluster "cognitive interaction technology" – cognition as a basis for natural interaction with technical systems. *it - Information Technology*, *51*(2), 112–118. doi: 10.1524/itit.2009.0532

Rosen, R. (2012). *Anticipatory systems*. Springer.

Rosenbaum, D. A. (2009). *Human motor control*. Academic press.

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & Van Der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human movement science*, *26*(4), 525–554.

Royakkers, L., & Steen, M. (2016). Developing tools to counteract and prevent suicide bomber incidents: A case study in value sensitive design. *Science and Engineering Ethics*, *23*(4), 1041–1058. doi: 10.1007/s11948-016-9832-8

Royce, W. W. (1987). Managing the development of large software systems: Concepts and techniques. In *Proceedings of the 9th international conference on software engineering* (pp. 328–338). Los Alamitos.

Said, N., Engelhart, M., Kirches, C., Körkel, S., & Holt, D. V. (2016). Applying mathematical optimization methods to an ACT-R instance-based learning model. *PLoS ONE*, *11*(7), 1-19. doi: 10.1371/journal.pone.0158832

Sand, O., Büttner, S., Paelke, V., & Röcker, C. (2016). smart.assembly – projection-based augmented reality for supporting assembly workers. In S. Lackey & R. Shumaker (Eds.), *Virtual, augmented and mixed reality* (pp. 643–652). Cham: Springer International Publishing.

Sauro, J. (2011, 2). *Measuring usability with the system usability scale (sus).* Retrieved 2020-05-25, from https://measuringu.com/sus/

Schack, T. (2004). The cognitive architecture of complex movement. *International journal of sport and exercise psychology*, *2*(4), 403–438.

Schack, T. (2012). Measuring mental representations. In G. Tenenbaum, R. C. Eklund, & A. Kamata (Eds.), *Measurement in sport and exercise psychology* (pp. 203–214). Champaign, IL: Human Kinetics.

Schack, T. (2020). Mental representation in action. In *Handbook of sport psychology* (p. 513-534). John Wiley & Sons, Ltd. doi: 10.1002/9781119568124.ch24

Schack, T., & Bar-Eli, M. (2007). Psychological factors in technical preparation. In B. Blumenstein, R. Lidor, & G. Tenenbaum (Eds.), *Psychology of sport training* (pp. 62–103). Oxford, UK: Meyer & Meyer Sport.

Schack, T., Bertollo, M., Koester, D., Maycock, J., & Essig, K. (2014). Technological advancements in sport psychology. In A. G. Papaioannou & D. Hackfort (Eds.), *Routledge companion to sport and exercise psychology : global perspectives and fundamental concepts* (pp. 953–965). Routledge.

Schack, T., Essig, K., Frank, C., & Koester, D. (2014). Mental representation and motor imagery training. *Frontiers in Human Neuroscience*, *8*, 328. doi: 10.3389/fnhum .2014.00328

Schack, T., & Hackfort, D. (2007). An action theory approach to applied sport psychology. *Handbook of sport psychology*, *3*, 332–351.

Schack, T., Hagan Jr., J. E., & Essig, K. (2020). New technologies in sport psychology practice. In D. Hackfort & R. J. Schinke (Eds.), *The routledge international encyclopedia of sport and exercise psychology. volume 2: Applied and practical measures* (p. 14). Routledge. doi: 10.4324/9781315187228

Schack, T., & Mechsner, F. (2006). Representation of motor skills in human long-term memory. *Neuroscience letters*, *391*(3), 77–81.

Schack, T., & Ritter, H. (2009). The cognitive nature of action—functional links between cognitive psychology, movement science, and robotics. *Progress in Brain Research*, *174*, 231–250.

Schack, T., & Ritter, H. (2013). Representation and learning in motor action–bridges between experimental research and cognitive robotics. *New ideas in psychology*, *31*(3), 258–269.

Schröder, M., & Ritter, H. (2017a). Deep learning for action recognition in augmented reality assistance systems. In *ACM SIGGRAPH 2017 Posters* (pp. 75:1 – 75:2).

Schröder, M., & Ritter, H. (2017b). Hand-object interaction detection with fully convolutional networks. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops* (pp. 18 – 25).

Schwaber, K. (1997). SCRUM development process. In J. Sutherland, C. Casanave, J. Miller, P. Patel, & G. Hollowell (Eds.), *Business object design and implementation* (pp. 117–134). New York: Springer. doi: 10.1007/978-1-4471-0947-1_11

Seegelke, C., & Schack, T. (2016). Cognitive representation of human action: theory, applications, and perspectives. *Frontiers in public health*, *4*.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). *Designing the user interface: Strategies for effective human-computer interaction* (6th ed.). Pearson.

Singh, M. (2008). U-SCRUM: An agile methodology for promoting usability. In *Agile 2008 conference*. IEEE. doi: 10.1109/agile.2008.33

Sinha, R. (2003). Persona development for information-rich domains. In *Chi '03 extended abstracts on human factors in computing systems* (pp. 830–831). New York: ACM. doi: 10.1145/765891.766017

Spiekermann, S. (2015). *Ethical IT innovation*. Auerbach Publications. doi: 10.1201/b19060

Stöckel, T., Hughes, C. M., & Schack, T. (2012). Representation of grasp postures and anticipatory motor planning in children. *Psychological research*, *76*(6), 768–776.

Strenge, B. (2013). *Integrationspotential von Ansätzen des Worth-Centred Development in agilen Softwareentwicklungsprozessen* (Master's thesis). Paderborn University, Germany.

Strenge, B., Vogel, L., & Schack, T. (2019). Computational assessment of long-term memory structures from SDA-M related to action sequences. *PLOS ONE*, *14*(2), 1-19. doi: 10.1371/journal.pone.0212414

Strotmeier, M. (2001). *Konzeption und prototypische Implementation eines Werkzeugs zur Unterstützung des Worth Centered Development-Ansatzes* (Diploma thesis). Paderborn University, Germany.

Subrahmanian, V., & Kumar, S. (2017). Predicting human behavior: The next frontiers. *Science*, *355*(6324), 489–489.

Sun, R. (2004). Desiderata for cognitive architectures. *Philosophical Psychology*, *17*(3), 341-373. doi: 10.1080/0951508042000286721

Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, *10*(2), 124–140. doi: 10.1016/j.cogsys.2008.07.002

Tang, A., Owen, C., Biocca, F., & Mou, W. (2003). Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 73–80). New York, NY, USA: ACM. doi: 10.1145/642611.642626

Tenenbaum, G., Hatfield, B. D., Eklund, R. C., Land, W. M., Calmeiro, L., Razon, S., & Schack, T. (2009). A conceptual framework for studying emotions-cognitions-performance linkage under conditions that vary in perceived pressure. In M. Raab, J. G. Johnson, & H. R. Heekeren (Eds.), *Mind and motion: The bidirectional link between thought and action* (Vol. 174, pp. 159–178). Elsevier. doi: 10.1016/S0079-6123(09)01314-4

Trafton, J. G., Hiatt, L. M., Harrison, A. M., Tamborello, F. P., Khemlani, S. S., & Schultz, A. C. (2013). Act-r/e: An embodied cognitive architecture for human-robot interaction. *J. Hum.-Robot Interact.*, *2*(1), 30–55. doi: 10.5898/JHRI.2.1.Trafton

Travel and Tourism Research Association (U.S.). (1996). *Annual conference* (No. 27). Bureau of Economic and Business Research, Graduate School of Business, University of Utah. Retrieved from `https://books.google.de/books?id=FUkXAQAAMAAJ`

Trudell, C., Hagiwara, Y., & Jie, M. (2014). *Humans replacing robots herald toyota's vision of future.* Retrieved 25 Nov 2019, from `https://www.bloomberg.com/news/articles/2014-04-06/humans-replacing-robots-herald-toyota-s-vision-of-future`

Tscherepanow, M., Kortkamp, M., Kühnel, S., Helbach, J., Schütz, C., & Schack, T. (2011). A feature selection approach for emulating the structure of mental representations. In B.-L. Lu, L. Zhang, & J. Kwok (Eds.), *Neural information processing: Lecture notes in computer science (lncs, volume 7064), 18th international conference iconip 2011, shanghai, china, november 13-17, 2011, proceedings, part iii* (pp. 639–648). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-24965-5_72

Umbrello, S. (2018). *Safe-(for whom?)-by-design: adopting a posthumanist ethics for technology design* (Master's thesis, York University). doi: 10.13140/RG.2.2.29726.38720

Umbrello, S. (2019). Imaginative value sensitive design: Using moral imagination theory to inform responsible technology design. *Science and Engineering Ethics*. doi: 10.1007/s11948-019-00104-4

Umbrello, S., & De Bellis, A. F. (2018). A value-sensitive design approach to intelligent agents. In R. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 395–409). CRC Press.

VanDeGrift, T., Dillon, H., & Camp, L. (2017). Changing the engineering student culture with respect to academic integrity and ethics. *Science and Engineering Ethics*, *23*(4), 1159–1182. doi: 10.1007/s11948-016-9823-9

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (Eds.). (2015). *Handbook of ethics, values, and technological design*. Springer Netherlands. doi: 10.1007/978-94-007-6970-0

Vogel, L., & Schack, T. (2016). The cognitive representation of complex actions in work processes: A technological approach for individual diagnostic in people with cognitive disabilities. *Journal of Sport & Exercise Psychology*, *38(Suppl.)*(Suppl.), 113.

Wachsmuth, I. (2008). Cognitive interaction technology: Humans, robots, and max. In *International conference on informatics education and research for knowledge-circulating society (icks 2008)* (p. 4-5). doi: 10.1109/ICKS.2008.34

Wachsmuth, S., Schulz, S., Lier, F., Siepmann, F., & Lütkebohle, I. (2012). The robot head "flobi": A research platform for cognitive interaction technology. In S. Wöfl (Ed.), *German conference on artificial intelligence, saarbrücken* (pp. 3–7). Deutsches Forschungszentrum für Künstliche Intelligenz.

Wang, X., Ong, S. K., & Nee, A. Y. C. (2016, Mar 01). A comprehensive survey of augmented reality assembly research. *Advances in Manufacturing*, *4*(1), 1–22. doi: 10.1007/s40436-015-0131-4

Weber, K. (2018). Extended model for ethical evaluation. In *Developing support technologies: Integrating multiple perspectives to create assistance that people really want* (pp. 257–263). Springer. doi: 10.1007/978-3-030-01836-8_25

Weigelt, M., Ahlmeyer, T., Lex, H., & Schack, T. (2011). The cognitive representation of a throwing technique in judo experts–technological ways for individual skill diagnostics in high-performance sports. *Psychology of Sport and Exercise*, *12*(3), 231–235.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In *Usability inspection methods* (p. 105–140). USA: John Wiley & Sons, Inc.

Whitefield, A., Wilson, F., & Dowell, J. (1991). A framework for human factors evaluation. *Behaviour & Information Technology*, *10*(1), 65–79. doi: 10.1080/01449299108924272

Wrede, S., Leichsenring, C., Holthaus, P., Hermann, T., Wachsmuth, S., & Team, T. C. (2017). The cognitive service robotics apartment: A versatile environment for human–machine interaction research. *KI - Künstliche Intelligenz*, *31*(3), 299–304. doi: 10.1007/s13218-017-0492-x

Yetim, F. (2011). Bringing discourse ethics to value sensitive design: Pathways toward a deliberative future. *AIS Transactions on Human-Computer Interaction*, *3*(2), 133–155. doi: 10.17705/1thci.00030

Zakani, F. R., Arhid, K., Bouksim, M., Gadi, T., & Aboulfatah, M. (2016). Kulczynski similarity index for objective evaluation of mesh segmentation algorithms. In *5th international conference on multimedia computing and systems (icmcs)* (p. 12-17). doi: 10.1109/ICMCS.2016.7905611

Zhu, Q., & Jesiek, B. K. (2017). A pragmatic approach to ethical decision-making in engineering practice: Characteristics, evaluation criteria, and implications for instruction and assessment. *Science and Engineering Ethics*, *23*(3), 663–679. doi: 10.1007/s11948-016-9826-6