On Cognitive Aspects of Human-Level Artificial Intelligence

Dissertation
zur Erlangung des Doktogrades
des Fachbereichs Humanwissenschaften
der Universität Osnabrück

vorgelegt
von

Dipl.-Math. Tarek R. Besold

aus
Bayreuth

Osnabrück, 2014

This is a cumulative doctoral thesis that includes articles published in peer-reviewed conference proceedings and books. The articles are not included in this electronic version of the dissertation due to copyright reasons. Instead, only the abstracts are embedded into this synopsis, but the full bibliographical reference and (where available) the corresponding URL of each article are provided.

Supervisor and first examiner: Prof. Dr. phil. Kai-Uwe Kühnberger
Co-examiner: Prof. Dr. rer. nat. Frank Jäkel
Co-examiner: Prof. Dr. rer. nat. Ute Schmid
Co-examiner: Dr. Iris van Rooij
Day of disputation: 19. December 2014

Abstract

Following an introduction to the context of Human-Level Artificial Intelligence (HLAI) and (computational) analogy research, a formal analysis assessing and qualifying the suitability of the Heuristic-Driven Theory Projection (HDTP) analogy-making framework for HLAI purposes is presented. An account of the application of HDTP (and analogy-based approaches in general) to the study and computational modeling of conceptual blending is outlined, before a proposal and initial proofs of concept for the application of computational analogy engines to modeling and analysis questions in education studies, teaching research, and the learning sciences are described.

Subsequently, the focus is changed from analogy-related aspects in learning and concept generation to rationality as another HLAI-relevant cognitive capacity. After outlining the relation between AI and rationality research, a new conceptual proposal for understanding and modeling rationality in a more human-adequate way is presented, together with a more specific analogy-centered account and an architectural sketch for the (re)implementation of certain aspects of rationality using HDTP.

The methods and formal framework used for the initial analysis of HDTP are then applied for proposing general guiding principles for models and approaches in HLAI, together with a proposal for a formal characterization grounding the notion of heuristics as used in cognitive and HLAI systems as additional application example.

Finally, work is reported trying to clarify the scientific status of HLAI and participating in the debate about (in)adequate means for assessing the progress of a computational system towards reaching (human-level) intelligence.

Two main objectives are achieved: Using analogy as starting point, examples are given as inductive evidence for how a cognitively-inspired approach to questions in HLAI can be fruitful by and within itself. Secondly, several advantages of this approach also with respect to overcoming certain intrinsic problems currently characterizing HLAI research in its entirety are exposed. Concerning individual outcomes, an analogy-based proposal for theory blending as special form of conceptual blending is exemplified; the usefulness of computational analogy frameworks for understanding learning and education is shown and a corresponding research program is suggested; a subject-centered notion of rationality and a sketch for how the resulting theory could computationally be modeled using an analogy framework is discussed; computational complexity and approximability considerations are introduced as guiding principles for work in HLAI; and the scientific status of HLAI, as well as two possible tests for assessing progress in HLAI, are addressed.

Contents

Acknowledgements

More than three and a half years passed between my first day as a PhD student and member of the AI research group at the Institute of Cognitive Science and the submission day of this thesis. During this time (and already before coming to Osnabrück) there were many people who supported and helped me in the context of my research and the projects which, among others, gave rise to the results documented in this thesis, but who also were with me in day-to-day life in and outside of academia. In the following, I want to thank a few of them – not because I have forgotten about all the others, but because the list simply would be too long.

First of all I am grateful to the members and students of the AI research group for their advice and will to collaborate on many of the topics mentioned in this thesis. Out of all of them I am especially indebted to Martin Möhrmann (whom I back then still got to know as Martin Schmidt) for his efforts in keeping HDTP running and developing the system further, for his support with invaluable software and hardware tips and tricks, and also simply for the occasional academic sanity check and for being a reliable friend.

Then there are the other members of the Cognitive Science PhD Programme who wrestled with their respective theses in parallel to me and my PhD work. It was great to compare and discuss their and my views, to be challenged and to get inspiration, but also support and affirmation, whenever needed. Thank you, Sascha Fink, Wendy Wilutzky, and all the others.

Thirdly, for a long (but still too short) period during these three and a half years there was a circle of IKW-related persons and personalities with very diverse backgrounds whom academia and cognitive science had brought together in Osnabrück, among them Carlos Zednik, Michael Baumgartner, and Constantin Rothkopf. I am grateful for having met all of you and for having been lucky enough to share more than one remarkable IKW Colloquium evening.

There was and is a number of people who inspired and deeply influenced me before and during my time at Osnabrück and whom by now (and for a long time to come) I hope to be able to count not only as colleagues, but also as friends:

- I owe a great debt of gratitude to my supervisor and head of the AI research group Kai-Uwe Kühnberger. Thank you for your close to unconditional support with all I did at the IKW and beyond, your advice and counsel regarding not only the topic

of my thesis but also academic life in general, and for sharing also many of your personal experiences and insights which were/are/will be relevant for my path as well.

- I am deeply indebted to Frank Jäkel, who always was and is willing to patiently discuss ideas and answer questions relating to almost every aspect of cognitive science (and far beyond that). Most things I learned about this quite particular discipline outside of my own area I acquired from him. I hope you will guard your enthusiasm and that also future generations of students will have the chance to experience it first hand.
- Michiel van Lambalgen, during my time at the University of Amsterdam, was the one who first talked to me about analogies, rationality, and reasoning, and who shaped my view on these topics, but (by his own way of being and thinking) also on scholarship and academia in general. I consider myself lucky having been given the privilege to study with you and experience all of this.

Clearly, as already stated initially, the list given up to here is far from exhaustive and important names are missing on it. Among the latter count, for instance, Helmar Gust, Iris van Rooij, Ute Schmid, Martin Stokhof, Maricarmen Martinez Baldares, Sara Uckelman, Krzysztof Apt, and many others.

And last but definitely not least there also were and are people outside of my little island within the academic ocean without whom this thesis would not have been possible, namely my family and friends. Thank you all for being with me, supporting me, believing in me, and also quite often simply reminding me that there is many more and at least equally important things outside of academia than I seemed to remember.

Whilst again this group is far too large to mention each and everyone, I am especially thankful to three people among them:

- My partner Erika who is and has been with me over the last two years at Osnabrück, reaffirming me when I was in doubt, providing me company when I felt alone, making me laugh when I was exhausted, and sharing the ups and downs the quite particular life of a PhD student and traveling academic involves. Thank you for standing by my side and walking this path with me!
- My parents Robert and Petra who always took and take an active interest and part in my life before and during my PhD studies. Although many aspects and internal procedures of academia must have seemed foreign to you and sometimes might have made you my deepest skeptics, you always gave me the feeling of being my most unconditional supporters and a safe haven if I should ever need one. Thank you for being there!

Affirmation

I hereby confirm that I wrote this thesis independently without the help of other parties and that I have not made use of resources other than those indicated. Data and concepts directly or indirectly imported from other sources have been marked indicating the source of origin. I guarantee that I significantly contributed to all materials used in this thesis (as reproduced in Part IV).
Further, this thesis has not been used in its present form or in a similar one to fulfill any other examination requirements neither within Germany nor abroad.

Osnabrück, 08. October 2014 ....

Abbreviations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
FOL	First-Order Logic
HDTP	Heuristic-Driven Theory Projection
HLAI	Human-Level Artificial Intelligence
PAI	Psychometric Artificial Intelligence
SMT	Structure Mapping Theory

Part I.

Synopsis

1. Opening Remarks

The present doctoral thesis includes work published in sixteen peer-reviewed conference or symposium contributions and book chapters which — although when seen individually might seem fairly heterogeneous — can all be subsumed under the overall headline of 'cognitive aspects of Human-Level Artificial Intelligence' research. This synopsis shall offer a concise but nonetheless accessible and comprehensive overview of the conducted work program and its main results. The individual chapters give a summary of the corresponding topical sub-projects, referring the interested reader to the respective publications in Part IV for details.

The following chapters present, after a short introduction to the overall context in Chap. 2, the outcomes of different interrelated sub-projects performed over the course of the last three years. Some of these investigations are fairly exploratory in nature, trying to identify approaches and to establish basic parts of conceptual frameworks addressing previously ignored topics (as, for example, the work on introducing computational analogy engines into the study of learning and education described in Chap. 5, or the proposals for more adequate and cognition-based perspectives on rationality in the context of Artificial Intelligence in Chap. 6). Others are situated more within the "classical" territory of Human-Level Artificial Intelligence (HLAI) research such as, for instance, the results presented in Chap. 3 dealing with aspects of the computational complexity and approximability of the Heuristic-Driven Theory Projection (HDTP) analogy engine or the considerations addressing the scientific status of HLAI and the quest for a proper test for (human-level) machine intelligence in Chap. 8

The line of narrative stringing together the individual parts follows a natural progression: After the introduction of the framing background of HLAI and (computational) analogy research in Chap. 2, an analysis assessing and qualifying HDTP's suitability for HLAI purposes is presented in Chap. 3. The following Chap. 4 then gives an account of a first (mostly exploratory) application scenario for HDTP (and analogy-based approaches in general) to an important question in HLAI, namely to the study and modeling of conceptual blending with computational means. A second applied sub-project, summarized in Chap. 5, describes a proposal and initial proofs of concept for the application of computational analogy engines to modeling and analysis questions in education and teaching research and the learning sciences. Chap. 6 subsequently changes the focus from analogy-related aspects in learning and concept generation to another HLAI-relevant cognitive capacity, namely rationality. After describing the relation between AI and rationality research in a quite general fashion, a new conceptual proposal for understanding and modeling rationality in a more human-adequate way is presented, together with a more specific analogy-centered account and an architectural sketch for the (re)implementation of certain aspects of rationality using HDTP. Chap. 7 then returns to the methods and formal framework used for the initial analysis of HDTP in Chap. 3 and shows how these can be relevant to HLAI research in general (also providing a second application example, namely a proposal for a formal characterization grounding the notion of heuristics as used in cognitive and HLAI systems). The penultimate Chap. 8 closes the circle in conceptually bridging back to the introduction to HLAI from the first chapter by reporting on work trying to clarify the scientific status of HLAI and participating in the debate about (in)adequate means for assessing the progress of a computational system towards reaching (human-level) intelligence. Chap. 9 finally gives some concluding remarks.

But before delving into the academic content of this thesis, a short clarification and/or disclaimer seems in place: Many of the questions addressed as part of the presented research effort are subject of ongoing study and (often heated) active debate within the respective scientific communities. In taking the particular approach presented throughout this thesis I do not want to claim that mine is the only possible or admissible perspective on the discussed issues, nor do I want to claim necessary superiority over alternative views. I rather want to show that the methods (and corresponding conceptual decisions) underlying my work can fruitfully be applied in the study of HLAI, emphasizing advantages by virtue of obtained positive results rather than by confrontation with (and subsequent dismissal of) competing conceptions.

2. Introduction to the Field(s)

Before addressing the individual lines of work unified in this thesis in Chap. 3 to 8, I want to give a short introduction to the two main pillars and connecting motives tying together the overall research project: HLAI as a quite specific research program, and computational analogy-making as one of the former's earliest, and ever since most active, core topics.

Therefore, in the following section I first give an orienting overview of HLAI as a field, providing the framing context and the background against which the work presented in subsequent chapters should be seen.[1] In the second section of the chapter, I offer a short introduction to the topic(s) of analogy and computational analogy-making, rooting work on computational models and systems for analogy-making in corresponding studies of analogy as a key mental function or facility from psychology and cognitive science. Overviews of technical notions and definitions relevant at different points throughout the thesis are given as appendix in Part II.

2.1. Human-Level Artificial Intelligence

When screening the literature from the early days of Artificial Intelligence (AI) research until today for a commonly agreed upon definition of what AI as a field of study is and what its aims are, it becomes very quickly obvious that there is no such thing as the desired universally hailed account of concept, means and goals. Still, in order to be able to introduce the notion of HLAI, standing at the core of this thesis, a definition of AI has to be fixed. Therefore, I adapt a phrasing originally to be found in the preface of Nilsson (2009): AI is that science devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment.

This understanding of AI is very inclusive, introducing a continuum of capacity levels ranging from fairly low-level technological systems and lower animals at the one end to humans (and possibly beyond) on the other end. HLAI is situated at the latter end of the just described spectrum, aiming at developing machines which can meaningfully be considered to be on par with humans in that they are similarly able to — among many others — reason, pursue and achieve goals, perceive and respond to different types of

[1] For a by far more detailed and complete description and assessment of past and present AI and HLAI research, see, for instance, Nilsson (2009).

stimuli from their environment (also, but not only, including language), process information, or engage in scientific and creative activities.[2] Still, neither AI nor HLAI by any means have to confine themselves to methods which are strictly biologically observable. Instead, for the time being, whatever technologically realizable means of (re)creating intelligence in an artificial system can be brought forward, have to and will be considered valid contributions to the research endeavor.

The history of HLAI goes back to the very origin of AI research. According to McCarthy (2007), the "first scientific discussion of human-level machine intelligence was apparently by Alan Turing" in a lecture to the London Mathematical Society in 1947, by this seemingly predating even Turing (1969)[3] and Turing (1950) (the latter of which is more commonly taken as the starting point of AI and HLAI history alike). A few years later, in his 1956 proposal for the Dartmouth Summer Study, John McCarthy famously laid out the program for generations of researchers to follow: "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.".[4]

And indeed, the spirit of McCarthy's just quoted statement still fuels each and every research endeavor in HLAI. Although there might be disagreement about the precise interpretation of "precisely described" or "simulate", and some researchers might want to expand the original phrasing in one way or another, HLAI as a field still rests on the assumption that the (re)creation of human-level intelligence by artificial means is possible and will eventually be achieved by scientific means. Clearly, over the decades different schools of thought left their traces also within HLAI research — such as, for example, the Physical Symbol System Hypothesis (Newell (1980)) and the logicist movement (Moore (1995)), the connectionist turn and the strong bias towards sub-symbolic forms of processing (Rumelhart et al. (1986); Smolensky (1987)), or the latest embodiment debates (Anderson (2003); Chrisley (2003)) — which has led to a certain diversification and independization between different streams within the overall field.

The present thesis situates itself within a subarea labeled "cognitive Human-Level Artificial Intelligence". This particular form of HLAI research is characterized by a strong inspirational and conceptual connection to results from cognitive science and psychology in approaching the task of a computational (re)creation of human-level mental faculties. Furthermore, in the present interpretation it clearly has to be considered to fall under the computational cognition paradigm as advocated, for instance, by Pylyshyn (1980). Also, all the models and systems presented and implemented in the context of this thesis are

[2] The line of research I refer to as HLAI is (in some cases with minor changes in particular research goals and approaches or overall ambition) also referred to as, among others, Artificial General Intelligence (AGI), "strong AI", "full AI", or "artificial cognition".

[3] Originally written as report for the National Physical Laboratory in 1948, but published only posthumously more than two decades later.

[4] A reproduction of the original proposal can be found as McCarthy et al. (2006).

symbolic in nature (i.e., logic-based). Still, this should not be taken as implicit renewal of the aforementioned Physical Symbol System Hypothesis in its unbending magnitude and with its extensive ramifications, but is exclusively due to the fact that for the purposes of the carried out studies a symbolic approach seemed more adequate.

2.2. Analogy and computational analogy-making

During the course of a day, as humans we use different kinds of reasoning processes: We solve puzzles, play instruments, or discuss problems. Often we will find ourselves in situations in which we apply our knowledge of a familiar situation to the structurally similar novel one. Today it is mostly undoubted that one of the basic elements of human cognition is the ability to see two a priori distinct domains as similar with respect to certain aspects, either based on their shared relational structure or (to a lesser extent) appearance — i.e., to recognize analogies and often use them for cross-domain transfer of knowledge and reasoning.[5] Some prominent cognitive scientists — with Hofstadter (2001) leading the way — even consider analogy the core of cognition itself.

As more cautiously described, for instance, by Schwering et al. (2009c), key abilities within everyday life, such as communication, social interaction, tool use and the handling of previously unseen situations crucially rely on the use of analogy-based strategies and procedures. Relational matching, one of the key mechanisms underlying analogy-making, is also one of the bases of perception, language, learning, memory and thinking, i.e., the constituent elements of most conceptions of cognition.[6]

Whilst analogies can be quite diverse in appearance and usage, it is widely agreed that on a procedural level at least three steps are indispensable (which also reappear in most, if not all computational models of analogy-making):

1. Retrieval: Given the target domain of an analogy (e.g., in form of a problem scenario a reasoner has to confront), the reasoner's memory has to be queried for similar cases encountered in the past. Candidate source domains have to be identified and made available to the analogy process.

2. Mapping: Given the target domain and a source domain (i.e., one of the candidate domains found during the retrieval step), respective domain elements which are hypothesized to stand in an analogical relation to each other have to be aligned. This process of pairing up domain elements can possibly give rise to insights about the internal structure of the domains, potentially also triggering domain-internal

[5] This definition implies a very broad conception of analogy, also covering phenomena such as similarity, metaphor and allegory (and, thus, can partially be seen in the Classical Greek tradition of analogy as shared abstraction described, among others, by Shelley (2003), and as related to the stance taken for example by Gentner and Markman (1997) for similarity and by Gentner et al. (1988) and Gentner et al. (2001) for metaphor).

[6] For an overview of psychological research on analogy see, for instance, Gentner and Smith (2013).

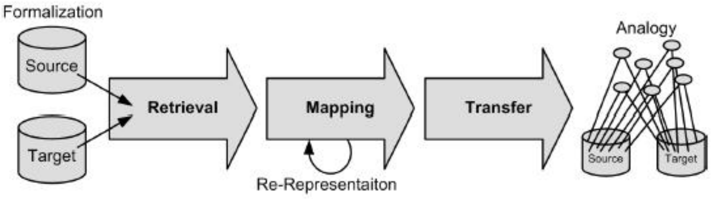

Figure 2.1.: A schematic overview of the standard conceptual approach to computational analogy-making from Schwering et al. (2009b).

restructuring and new conceptualizations. (In the case of computational analogy-making, the alignment is mostly based on structural and syntactic properties of the respective representation formalisms applied in the domains, and the process of restructuring the domains is called re-representation. See also Fig. 2.1.)

3. Transfer/evaluation: Once a mapping has been established between source and target domain, knowledge can be transferred from the (better informed) source to the (more sparse) target domain using the alignments from the mapping phase as guidance for the potentially needed knowledge adaptation during the cross-domain transfer. Once the target domain has been enriched, a final step of evaluation (possibly also involving reasoning within and, once again, restructuring of the target domain) judges the established analogy. This judgement can then also be used to decide whether the analogy-making was successful, or if another candidate domain should be considered instead of the used source domain — either by returning to the mapping phase and using another candidate from the collection of retrieved cases, or even by returning to the retrieval phase and (taking into account the insights about the target domain gained during restructuring and transfer) starting the entire procedure anew.

Because of the described crucial role of analogy in human cognition researchers on the computational side of cognitive science and in AI also very quickly got interested in the topic and have been creating computational models of analogy-making basically since the advent of computer systems. Although the developed models and implemented systems differ vastly in their precise specifications and computational paradigms (some being symbolic, some connectionist, others hybrid), on the level of procedural abstraction most also adhere to the just outlined retrieval–mapping–transfer/evaluation triad and can be conceptualized as shown in Fig. 2.1.

The resulting history of computational analogy systems starts with Reitman et al. (1964)'s ARGUS and Evans (1964)'s ANALOGY in the late 1950s and early 1960s, contains, for instance, Winston (1980)'s work on analogy and learning, and features systems as prominent as Hofstadter and Mitchell (1994)'s Copycat or Falkenhainer et al. (1989)'s famous Structure-Mapping Engine and Gentner and Forbus (1991)'s MAC/FAC.[7]

[7] For an informative overview of different architectural and conceptual paradigms for computational analogy engines and of well-known implemented systems see, for example, Besold (2011).

3. Computational Properties of Analogy-Making Using the Heuristic-Driven Theory Projection Framework

Whilst Falkenhainer et al. (1989)'s Structure-Mapping Engine and Gentner and Forbus (1991)'s MAC/FAC, mentioned at the end of the previous chapter, implement a version of Gentner (1983)'s Structure Mapping Theory (SMT), more recently a different, generalization-based approach has been proposed: HDTP (Schmidt et al. (2014)).

Both paradigms, SMT and HDTP, are very similar in that they are symbolic (i.e., operating on domain theories expressed in logic-based languages) and during the mapping stage heavily rely on syntactical properties of the respective representation languages and domain formalizations for pairing up domain elements. Still, whilst SMT explicitly proclaims that the mapping between domains is established directly from elements in the source domain to elements in the target domain, and that the subsequent transfer/evaluation step is exclusively guided by groupings of these individual correspondences, HDTP explicitly computes a generalization of the source and target domain theories into a least general subsuming theory which later determines the transfer/evaluation phase.[1] The process of analogy-making in HDTP can be conceptualized as shown in Fig. 3.1.

HDTP aims at being a mathematically sound framework for the computation of analogical relations and inferences between domains which are given in form of a many-sorted first order logic representation. Source and target domain are handed over to the system in terms of finite axiomatizations and HDTP tries to align pairs of formulae from the two domains by means of restricted higher-order anti-unification as introduced by Krumnack et al. (2007): Given two terms, one from each domain, HDTP computes an anti-instance in which distinct subterms have been replaced by variables so that the anti-instance can be seen as a meaningful generalization of the input terms. As already indicated by the name, the class of admissible substitution operations is limited. On each expression, only renamings, fixations, argument insertions, and permutations may be performed. By this

[1] Here, subsumption has to be understood in the following sense: The joint generalized theory subsumes the original source theory and target theory in that each of the latter can be re-obtained by applying certain substitution operations to the generalization (see Part II, Chap. A for details). In this way the joint generalization encompasses both domain theories at a time as more specific variants.

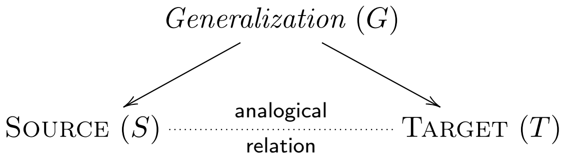

Figure 3.1.: A schematic overview of HDTP's generalization-based approach to analogy.

process, HDTP tries to find the least general generalization of the input terms, which (due to the higher-order nature of the anti-unification) is not unique. In order to solve this problem, current implementations of HDTP rank possible generalizations according to a complexity measure on the chain of substitutions — the respective values of which are taken as heuristic costs — and returns the least expensive solution as preferred one. HDTP extends the notion of generalization from terms to formulae by basically treating formulae in clause form and terms alike. Finally, as analogies rarely rely exclusively on one isolated pair of formulae from source and target domain, but usually encompass sets of formulae (possibly completely covering one or even both input domains), a process iteratively selecting pairs of formulae for generalization has been included. The selection of formulae is again based on a heuristic component: Mappings in which substitutions can be reused get assigned a lower cost than isolated substitutions, leading to a preference for coherent over incoherent mappings.[2]

Concerning applications of HDTP, for instance Guhe et al. (2010) apply the framework in modeling a potential inductive analogy-based process for establishing the fundamental concepts of arithmetics, Guhe et al. (2011) describe an account of a process blending different conceptualizations of number into new conceptualizations recognizing common features and combining distinct ones, and Schwering et al. (2009a) use HDTP for solving geometric analogies. Further successful and completely worked out application examples, among others, also can be found as part of this thesis in Chap. 5 on modeling analogy-related processes from education and learning. Synoptically, these and related studies illustrate the overall potential of HDTP in terms of expressivity and domain generality, and establish its suitability for application to also fairly complex analogy-making tasks.

Complementing these efforts and completing the picture, Robere and Besold (2012) and Besold and Robere (2013) provide an investigation into the computational complexity and approximability of the approach.[3] Viewing HDTP from the perspective of theoretical computer science as a tool for the computation of generalizations between domain theories by means of a special form of anti-unification (i.e., putting it under the umbrella of

[2] For a short overview of further formal aspects of HDTP (and restricted higher-order anti-unification in particular) see Part II, Chap. A. For details specifically concerning the heuristic aspects of HDTP see, for instance, Schwering et al. (2009b) and Schmidt et al. (2011).

[3] See Part IV for full versions of both papers.

rewriting research), properties, such as computational complexity and approximability, will not necessarily play a central role in the evaluation of the overall approach and system. But taking a cognitive HLAI view, the picture changes: Given that the premise is to ultimately (re)create human-level intelligence with computational means, properties, such as the amount of resources required for computation, become critical factors.[4]

Robere and Besold (2012) present a parameterized complexity analysis of HDTP in its present form using restricted higher-order anti-unification, and also of an earlier approach based on reducing higher-order anti-unification to first-order anti-unification as described, for example, by Gust et al. (2006).[5] The given analysis focuses exclusively on the analogical matching of input theories and leaves out the (potentially also required) re-representation of input theories by deduction in First-Order Logic (FOL). Although HDTP in its entirety encompasses both parts, from a complexity point of view re-representation can be identified fairly straightforwardly as undecidable due to the undecidability of FOL.

By defining three increasingly complex and expressive versions of higher-order anti-unification — successively admitting additional types of unit substitutions to be included in the anti-unification process — Robere and Besold (2012) obtain the following results for analogical matching using restricted higher-order anti-unification:[6]

1. Anti-unification using only renamings and fixations is solvable in polynomial time.

2. Anti-unification using renamings, fixations and a bounded number of permutations is NP-complete and $\mathsf{W}[1]$-hard with respect to the minimum number of higher arity variables and the maximum number of permutations, and becomes fixed-parameter tractable only with respect to the maximum arity and the maximum number of subterms of the input terms, together with the maximum number of permutations.

3. Anti-unification using renamings, fixations, a bounded number of permutations, and argument insertions is NP-complete and $\mathsf{W}[1]$-hard with respect to the minimum number of higher arity variables, the maximum number of permutations and the maximum number of argument insertions.

Additionally, concerning the earlier version of HDTP based on reducing certain higher-order to first-order anti-unifications by the introduction of subterms built from so called *admissible sequences*, Robere and Besold (2012) show a similar result, establishing NP-completeness and $\mathsf{W}[2]$-hardness of the analogical matching (the latter with respect to a key parameter inherent to the admissible sequence-based approach).

Besold and Robere (2013) then complement the parameterized complexity analysis of HDTP by an assessment of the approximation theoretic complexity of the analogical

[4] See Chap. 7 for a more detailed discussion of relevant arguments.

[5] See Part II, Chap. B for a short overview of required basic notions from parameterized complexity theory.

[6] In the following, higher arity variables are variables of arity at least 1.

matching.[7] Starting out from the just reproduced results, using a measure of complexity for any composition of the substitutions allowed in restricted higher-order anti-unification introduced by Krumnack et al. (2007), and defining the optimization problem as trying to find a generalization which maximizes the complexity over all generalizations — i.e., a generalization which maximizes the "information load" (Krumnack et al. (2007)) over all chosen generalizations —, it is shown that the problem of analogical matching in the restricted higher-order anti-unification setting using renamings, fixations, and a bounded number of permutations does not allow for constant-factor approximation algorithms.

Taking the parameterized and the approximation complexity results together, even when presuming that a solution for the undecidability problem for re-representation caused by the choice for full many-sorted FOL as language would be found, the derived (mostly negative) results concerning the analogical matching cast a shadow over the ambition of using HDTP as basis for general computational theories of high-level cognitive capacities with a certain degree of cognitive plausibility (as, for example, suggested by relating HDTP to the proposal of their integrated cognitive architecture I-Cog by Kühnberger et al. (2007) and Kühnberger et al. (2008), or still hinted at by Martinez et al. (2012) when proposing to use HDTP as basis for modeling creativity and productivity issues in artificial systems): Severe scalability problems in terms of needed resources for computation are to be expected when dealing with anything but small and strongly restricted example scenarios. On the other hand, mostly due to the more refined perspective parameterized complexity theory offers over classical forms of analysis, the results point to specific elements of the approach as sources for intractability (such as, for instance, the use of permutations for restructuring formulae) and the approximation analysis adds additional weight and precision to the findings (by showing that, even when trying to solve the analogical matching including the use of permutations only approximately, no significant change in the state of affairs can be expected).

Also, the discussed negative complexity and approximability properties do not devalue HDTP's approach in general, they simply contribute to clarifying its characteristics and to specifying possible application scenarios — in doing so helping to avoid potentially time- and resource-intense (mis)developments with little chance of success. And whilst most likely not being suitable as basis for (parts of) a real-time reactive HLAI system, HDTP's perspective on modeling and implementing analogy-making and related cognitive capacities can still fruitfully be applied to create computational accounts useful and suitable for clarifying the role of analogy-related faculties in conceptual blending (see, for example, Chap. 4), learning and education (see, for instance, Chap. 5), rationality and decision-making (see, for example, Chap. 6), and many other domains.

From a conceptual perspective, the overall approach and the types of analyses applied seamlessly tie into the bigger framework of resource-sensitive modeling and computa-

[7] See Part II, Chap. C for a short overview of required basic notions from approximation theory.

tion in HLAI discussed in Chap. 7, providing an application case study and a proof of feasibility for the theoretical guidelines outlined there.

4. Theory Blending Using the Heuristic-Driven Theory Projection Framework

Boden (2003) identifies three forms of creativity: exploratory, transformational, and combinatorial. The label exploratory refers to creativity which arises from a thorough and persistent search of a well-understood domain (i.e., within an already established conceptual space), whilst transformational creativity either involves the removal of constraints and limitations from the initial domain definition, or the rejection of characteristic assumptions forming part of the specification of the creative problem (or both). Combinatorial creativity shares traits of both other forms in that it arises from a combinatorial process joining familiar ideas (in the form of, for instance, concepts, theories, or artworks) in an unfamiliar way, by this producing novel ideas.

Computationally modeling the latter form of creativity turns out to be surprisingly complicated: Although the overall idea of combining preexisting ideas into new ones seems fairly intuitive and straightforward, when looking at it from a more formal perspective at the current stage neither can a precise algorithmic characterization be given, nor are at least the details of a possible computational-level theory describing the process(es) at work well understood. Still, in recent years a proposal by Fauconnier and Turner (1998) called conceptual blending (or conceptual integration) has influenced and reinvigorated studies trying to unravel the general cognitive principles operating during creative thought. In their theory, conceptual blending constitutes a subconscious process which allows for the combination of certain elements (and their relations) from originally distinct conceptual spaces into a new unified space combining these previously separate elements and allowing to perform reasoning and inference over the combination.

Unfortunately, Fauconnier and Turner (1998), and also Fauconnier and Turner (2002) and Fauconnier and Turner (2008), remain mostly silent concerning details needed for a proper computational modeling of conceptual blending as cognitive capacity — neither do they provide a fully worked out and formalized theory themselves, nor does their informal account capture key properties and functionalities as, for example, the retrieval of input spaces, the selection and transfer of elements from the input into the blend space, or the further combination of possibly mutually contradictory elements in the blend. In short: The theory does not specify how the blending process is supposed to work.

These shortcomings notwithstanding, several researchers in AI and computational cognitive modeling have used the provided conceptual descriptions as starting point for proposing possible refinements and implementations. Goguen and Harrell (2010) propose a conceptual blending-based approach to the analysis of the style of multimedia content in terms of blending principles and also provide an experimental implementation, Pereira (2007) tries to develop a computationally plausible model of several hypothesized sub-parts of conceptual blending, Thagard and Stewart (2011) exemplify how creative thinking could arise from using convolution to combine neural patterns into ones which are potentially novel and useful, and Veale and O'Donoghue (2000) present their computational model of conceptual integration and propose several extensions to the at the time actual view on conceptual blending.

It is this tradition of trying to fill in the computational gaps in the original theoretical accounts of conceptual blending — in order to subsequently attempt to use the resulting framework for eventually implementing productively creative systems — in which also the work presented by Martinez et al. (2011) and Martinez et al. (2012) has to be seen.[1] The former paper proposes a logic-based framework for blending and metaphor-making building on HDTP and expanding its analogy-centered account to also encompass theory blending (as more specific form of conceptual blending), points out how different relevant forms of cross-domain reasoning (i.e., analogy, cross-domain generalization, cross-domain specialization, and the detection of congruence relations) can be implemented using HDTP, and prototypically sketches different application case studies of the resulting framework, among others, to blending in mathematics or the solving of rationality puzzles. Martinez et al. (2012) subsequently revisit their former work, provide additional theoretical considerations on theory blending, engage in more detail with the modeling of Argand (1813)'s discovery of the complex plane as result of a conceptual blending process, and in conclusion provide an outlook on future applications of conceptual blending in modeling creative capacities in next generation AI systems.

The interpretation of conceptual blending applied by Martinez et al. (2011) and Martinez et al. (2012) is based on the account given by Goguen (2006): Given two domain theories I_1 and I_2 representing two conceptualizations, first compute a generalization G and then construct the blend space B in such a way as to preserve the correlations between I_1 and I_2 given by G (see also Fig. 4.1).

In this view, the morphisms mapping the axioms of one domain theory to another are induced by signature morphisms over the symbols of the representation languages — an analogical correspondence is assumed to exist between symbols in I_1 and I_2 coming from the same symbol in G. As incompatibilities might pertain between the domains, the morphisms from I_1 and I_2 to B are possibly only partial (i.e., not all axioms from the domain theories are mapped to the blend). In Goguen (2006)'s category theory-based framework, B is the smallest theory comprising as much as possible from I_1 and I_2 while reflecting the commonalities of I_1 and I_2 encoded in the generalization G.

[1] See Part IV for full versions of both papers.

$$\begin{tikzcd}
 & B & \\
I_1 \arrow[ur] & & I_2 \arrow[ul] \\
 & G \arrow[ul] \arrow[ur] \arrow[uu] &
\end{tikzcd}$$

Figure 4.1.: A conceptual overview of Goguen (2006)'s account of conceptual blending.

This approach clearly offers itself to a (re)conceptualization and (re)implementation using HDTP: Whilst intra-domain reasoning can be performed with classical logic calculi over the many-sorted FOL language, the computation of a generalization G from two input domains I_1 and I_2 (involving cross-domain generalization, cross-domain specialization, and possibly the detection of congruence relations as used, for instance, by Guhe et al. (2011)) is one of HDTP's core functionalities. Thus, basically the entire lower half of Fig. 4.1 is naturally covered by the standard mechanisms of HDTP (also compare Fig. 4.1 with Fig. 3.1).[2] The only additional element needed is a mechanism for using the information provided by the domain theories I_1 and I_2 together with the generalization G for computing the blend B.

Martinez et al. (2011) and Martinez et al. (2012) use this conceptual framework for reconstructing a famous occurrence of conceptual blending in the history of mathematics: The discovery of the geometrical interpretation of the complex numbers as complex plane by Argand (1813). As reproduced in Stedall (2008), Argand used a succession of consecutive blending steps (partially involving preexisting domain theories, partially using the results from earlier blending steps in the continuation of the process) for deriving his theory. The resulting network of theories can be conceptualized as shown in Fig. 4.2.

In the same vein, Martinez et al. (2011) (and in less detail also Martinez et al. (2012)) offer an explanation of Tversky and Kahneman (1983)'s famous Linda problem which seems to be an instance of the so-called conjunction fallacy: Subjects are told a story about a bank teller named Linda and are subsequently asked to order different statements about Linda according to their perceived probability given the previously acquired information from the story. Quite reliably, subjects tend to consider a statement combining two properties into one as more likely as the individual property statements, by this contradicting the basic laws of probability theory in that the joint probability of two events has to be less or at most equal to the probability of each individual event. But instead of following Tversky and Kahneman (1983) in their attempt at an explanation of this phenomenon relying on the introduction of the notion of representativeness as

[2] For the purpose of conceptual blending, the distinction between a source domain and a target domain as in the case of computational analogy-making becomes obsolete.

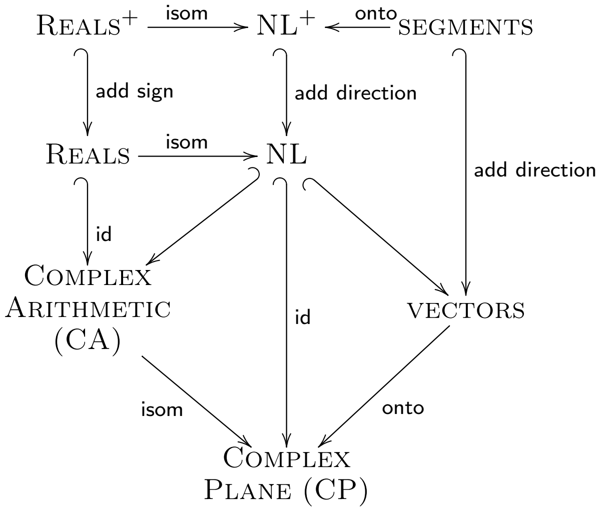

Figure 4.2.: The network of domains involved in Argand (1813)'s article as reconstructed by Martinez et al. (2011) and Martinez et al. (2012): Argand's addition to the — in his time already preexisting — network of interconnected mathematical concepts and theories are the nodes vectors for the vector domain and Complex Plane (CP), which together with Complex Arithmetic (CA) and the number line NL form the shape of a blending diagram (also see Fig. 4.1). In the diagram, arrows indicate the direction of (partial) theory morphisms, labels indicate constraints at the level of models, and curved tails replace "injective" labels on the arrows.

driving force behind this deviation from (probabilistic) rationality, Martinez et al. (2011) provide formalizations of relevant parts of the story about Linda and of the properties involved in the alleged violation of rational behavior and use these to subsequently outline how the HDTP-based approach to conceptual blending in combination with some of the principles and dynamics internal to HDTP can provide an account (re)rationalizing the subjects' answering behavior.[3]

In summary, these and related studies, such as, for example, the work reported by Abdel-Fattah et al. (2012), Abdel-Fattah and Krumnack (2013) and Abdel-Fattah et al. (2013), give proof of the feasibility of an analogy-based approach to conceptual blending and the applicability of HDTP as modeling tool to occurrences of conceptual blending in diverse contexts, ranging from the described reconstruction of mathematical discoveries and explanations for rationality puzzles, through the assignment of meaning to previously unknown compound nouns, to the interpretation of counterfactual statements.

[3] Also see Chap. 6 for a proposal of why the obvious violation of basic laws of probability theory by the subjects does not necessarily have to constitute an instance of irrational behavior or qualify the subjects as irrational in general.

5. Computational Analogy-Making in Education

As already stated in Chap. 2, analogy and related faculties (as, for instance, conceptual blending) are pervasive in our daily life and form part of many important high-level human cognitive functions and abilities. One of the most crucial among the latter is the human capacity to learn, i.e., to acquire new knowledge and to develop previously unmastered capabilities based on either unsupervised and uninstructed observation and experimentation or through guided and targeted educational processes: Gentner et al. (2001) proposed analogy as an essential part of the human ability to learn abstract concepts, and Ross (1987) showed its role in learning new procedures. Also, according to Novick (1988), analogy seems to play a crucial role in relation to the ability to transfer representations across contexts or to adapt to novel contexts (see, for example, Holyoak and Thagard (1995)), and Goswami (2001) found a strong influence of analogy within children's cognitive repertoire for learning about the world.

Of course, based on these psychological findings, analogy also started to receive attention in education and the learning sciences: Duit (1991) provides evidence for the role of analogy in facilitating learners' construction process of new ideas and conceptions on the grounds of already available concepts, and Arnold and Millar (1996) show that analogies can foster understanding of scientific explanations. On the other hand, analogy is not a cure-all as unsuccessful analogies may produce misunderstandings and can possibly result in harmful misconceptions (see, for example, the work by Clement (1993) and Zook and Maier (1994)). All these findings notwithstanding, following among others Akgul (2006), the current level of knowledge about analogy as an instructional device in everyday practice has to be considered quite low.

In a series of papers building upon each other, Besold (2013a), Besold et al. (2013), Besold and Kühnberger (2014) and Besold (2014b) model and analyze with computational means the use of analogy-based methods in physics and mathematics teaching scenarios taken from a primary school classroom setting, providing evidence for the general suitability of computational analogy engines for usage in studying methods and tools applied in teaching and education.[1] These efforts form part of a bigger program aiming to introduce computational approaches and models of analogy also to education and the learning sciences, by this on the one hand enriching the methodological repertoire available to these fields of study and on the other hand opening up new domains to AI and

[1] See Part IV for full versions of the referenced papers.

cognitive systems research.

Besold (2013a) proposes three (partially overlapping) possible scenarios of use for computational analogy systems in education and teaching-related topics:

- **Modeling and analysis:** Symbol-based analogy engines can be used for explicitly modeling and understanding the conceptual mode of operation of analogies in a teaching context, addressing a level of detail situated between the level of computational theory and the level of representation and algorithm in Marr (1982)'s Tri-Level hierarchy.

- **Exploration and testing:** Frameworks allowing for an incorporation of parameters from experimental and classroom data, mirroring children's cognitive capacities and limitations in situation-adequate ways, into the mechanisms of the analogy engine may be used for exploring, developing and testing given analogies for a teaching context. Also, in a way simulating children's analogy generation and understanding in a specific situation, the system could be used for discovering the analogies which the children might find between two domains a teacher wants to use in a selected educational context.

- **Discovery and guidance:** Given two domains by the teacher, the analogy engine can be used for discovering what analogies generally could arise between these domains and how the analogy-making process might have to be guided (e.g., what additional framing facts have to be included in the initial domain theories) to obtain one specific, previously planned analogy (or set of analogies) as result of the process.

As proof of concept for the proposed approach, Besold (2013a) provides a reconstruction of the string circuit analogy for electric current in form of an HDTP-based model. This particular analogy is used in science classes for 8 to 9 year old children and aims at facilitating the understanding of the basic principles of electric current: Participants are placed in a circle and asked to loosely support a string loop which one person (normally the teacher) subsequently makes circulate. Using prior knowledge about electric circuits, obtained when playing with simple circuits, students then are asked to conjecture how the person moving the string relates to the battery in an electric circuit, how the string loop relates to the electric current, etc. The HDTP model reproduces findings concerning the alignment of domain elements from a study by Guerra-Ramos (2011) and additionally clarifies the involved conceptual transfers and necessary elements of prior knowledge in each domain.

Besold et al. (2013) then provide a second, more complex example for the reconstruction of an analogy-based teaching method. There, Schwank et al. (2005)'s Calculation Circular Staircase, a teaching tool for basic arithmetics and the interpretation of the natural numbers as ordinal numbers used with children in primary school, is modeled using

HDTP. The staircase allows children to playfully develop an understanding of addition and subtraction within the basic number space of the naturals through repeated constructive transformation processes which are grounded in motor-based interaction with the tool. In this case, the HDTP model sheds light on how the interaction between the structure and the built-in principles of the Calculation Circular Staircase on the one hand, and the memorized declarative counting procedures learned in school on the other hand can bring forth the targeted conception of the space of the natural numbers together with the basic arithmetic processes operating on it.

Another mathematics teaching tool — in several ways similar to the Calculation Circular Staircase — is Schnalle and Schwank (2006)'s Number Highrise. Besold and Kühnberger (2014) use HDTP to construct a model of this mathematical toy world designed for exploring the natural number space up to 100 and for discovering multiplication-based relations within this space. Analogous to the Calculation Circular Staircase, children through (partially guided, partially unsupervised) playful interaction with the teaching device discover numerical relationships within the range of 1 to 100, such as, for instance, the concept of the least common multiple and the greatest common factor, or the notion of prime numbers. The provided computational model again explicates the underlying structural relations and governing laws and offers an account of how the construction-based experiences and insights from play, in combination with previously memorized facts about natural numbers and times tables, can give rise to the discovery and understanding of more complex notions.

Besold (2014b) concludes the cycle of papers by summarizing the initial conceptual ideas of Besold (2013a), together with accounts of the HDTP models of the string circuit analogy, the Calculation Circular Staircase, and the Number Highrise, and embedding the entire program into the framework of research on the role of sensorimotor experiences in learning and teaching on the one hand, and the computational study of creativity and cognitive capacities, such as learning and concept formation, on the other hand.

In general, the examples presented in the different studies provide evidence for the feasibility and usefulness of applying computational methods and, more precisely, computational analogy engines to research questions in education and the learning sciences. But this clearly is not a one-way transfer: The insights resulting from this type of work are not only useful for researchers on the psychological or pedagogical side of the interaction, but they can also be useful for understanding how to model and (re)create cognitive capacities in AI systems. By modeling children's analogy-based learning and development processes, general insights about the mechanisms at work can be gained which then can again fruitfully be applied in the construction of new and the improvement of already existing computational frameworks and paradigms.

With respect to related studies, although the discussed application of computational analogy frameworks to the modeling and analysis of analogy-based teaching tools and techniques is currently not pursued in a systematic way outside the presented project, this line of work continues efforts conducted, for instance, by Thagard et al. (1989) and by Siegler (1989): Whilst the former present a theory and implementation of analogical mapping in the context of explanations of unfamiliar phenomena, such as, for example, used by chemistry teachers, the latter conjectures how Falkenhainer et al. (1989)'s Structure-Mapping Engine could be put to use in gaining insights about developmental aspects of analogy use.

6. Rationality, Analogy, and Human-Level Artificial Intelligence

One of the qualities allegedly setting apart humans from their fellow animals is the former ascribed capacity to think rationally and act accordingly. Already Aristotle characterized humans as rational animals ("zoon logikon") in his Metaphysics (see, for example, Tredennick (1933)) and as being imbued by a rational principle in his Nicomachean Ethics (see, for instance, Broadie and Rowe (2002)). These and similar considerations by other philosophers and scholars over the centuries fueled a continuous stream of investigations into the nature and properties of (human) rationality. Over the course of these investigations, rationality got partially uncoupled from its human hosts, transforming the notion from a mainly descriptive account of a reasoner's behavior (with the reasoner being assumed to be mostly rational due to her nature as rational animal) to a normative and prescriptive theory of how humans should reason in theory and practice.

It should therefore not come by surprise that currently no unanimously accepted definition of rationality exists, but that there are several competing frameworks for modeling rationality (and establishing a normative theory). Following Besold (2013d), a non-exhaustive list of types of models has to contain at least four main categories: Logic-based models (see, for example, Evans (2002)), probability-based models (see, for instance, Griffiths et al. (2008)), game-theoretically based models (see, for example, Osborne and Rubinstein (1994)), and heuristic-based models (see, for instance, Gigerenzer et al. (2011)). The respective definitions of rationality (and the corresponding frameworks) are in many cases almost orthogonal to each other, when taken together making them at best incommensurable, if not inconsistent or even partly contradictory. Also, in many cases the predictive power turns out to be limited as the focus lies on the normative or postdictive-explanatory aspects of the respective models.

But independently of whether rationality is taken as descriptive, normative/prescriptive or, possibly, predictive characterization of a reasoner's mental processes and behavior, of course AI research and (even more) specifically work in HLAI from the very beginning has been interested in modeling rationality and rational behavior with computational means: If an artificial system were to be considered intelligent and human-like in its capabilities and actions, deviations from rationality and rational behavior in cases where humans seem to adhere to these standards would be judged as major setbacks. Also, if an artificial system were to interact closely with humans, a feasible notion of rationality would most likely be required within the system to allow for predictions of human behavior by the AI and vice versa.

The first section of this chapter, based on Besold (2013d), accordingly shortly discusses the history and status quo of rationality as topic of study within AI and HLAI and then continues to advocate dedicated research on (artificial) rationality, aiming at developing accounts of rationality suitable for the specific needs and purposes of (HL)AI. The second section, following Besold and Kühnberger (2012), subsequently introduces a subject-centered notion of rationality as alternative proposal to existing mainstream theories of rationality and, summarizing work presented in Besold (2013c), sketches how such a notion could be conceptualized and implemented using models of analogy and HDTP as computational analogy engine.[1]

6.1. Rationality and Artificial Intelligence

As also pointed out, for instance, by Russell (1997), the concept of rational agency has for a long time been considered a promising candidate when trying to define the notion of intelligence AI is striving for. And although there are several different notions of how rationality could precisely be interpreted or instantiated, for the purpose of providing a short sketch of the role(s) rationality played and could play in AI research, Besold (2013d) uses a quite general conception only demanding for an action/behavior to adhere to the "principle of rationality" to be considered as rational: Solutions to a task have to constitute the least effortful, direct course of action to attain as much of a goal as it is possible given current environmental and other constraints.[2]

As already advocated previously, AI and, even more, HLAI research should be interested in rationality for at least two reasons: On the one hand, when aiming for human-like intelligence in an artificial system, modeling rationality on a human level has to be considered a milestone towards the overall goal. On the other hand, having a feasible understanding and account of rationality in an AI system can play a key role in interaction with humans, making the interaction between natural and artificial agents more natural and seamless. As, for example, Doyle (1992) points out, there is reason to believe that a theory of rationality might in general at some point equal mathematical logic in its importance for mechanizing reasoning.

Still, when having a closer look at the status quo of rationality research within AI and HLAI, it has to be noted that the employed notions of rationality have stayed fairly close to their fields of origin and have mostly been taken over without proper adaption to the chances and limitations introduced by the computational AI/HLAI context: In order to just mention a few shortcomings, neither has the explicit access to computational properties of processes widely been taken into account, nor have more prominent attempts

[1] See Part IV for full versions of the referenced papers.

[2] This understanding of rationality also covers, among others, the different forms of rationality considered by Russell (1997).

been undertaken to exploit synergies with previous theoretical or applied work as, for example, to be found in theoretical computer science or cognitive modeling and cognitive systems research.

This state of affairs has to be considered even more surprising when having a look at the advantages of a targeted and specifically AI-rooted approach to rationality research would offer: A more holistic perspective on rationality within artificial systems could be achieved, combining existing theories and tools from all over AI, computational factors (such as, for instance, computational complexity) could explicitly be addressed and widely be taken into account, and the nature of AI as partially engineering-based endeavor (resulting in the creation of new and stepwise improvement of existing runnable systems) would allow for empirical testing and development of theories in a cycle of theory-practice-coevolution. Finally, it has to be noted that such an approach in all likelihood would not only profit AI and HLAI as a field of study, but could also have beneficial reflections back on more classical fields of rationality research, such as, psychology, economics, and philosophy, as the latter could be provided with testbeds for their models and theories, whilst also getting inspiration from approaches and perspectives previously limited to within AI.

6.2. An analogy-based approach to rationality in Human-Level Artificial Intelligence

Besold (2013d) also provides a first sketch of a concept further theoretically developed by Besold and Kühnberger (2012) and considered in an implementation context by Besold (2013c): A notion of "subject-centered rationality" with a strong emphasis on incorporating reasoner-specific constraints and possibilities and a focus on the positive and predictive aspects of the resulting models, sharing important elements, for example, with Simon (1956)'s "bounded rationality" or with the subjective concept of rationality by Gilboa and Schmeidler (2001).

The advocated position takes its roots in the interpretation of rationality as a genuinely human faculty (following the already mentioned understanding of man as rational animal) and, thus, as a notion empirically defined by human reasoners — which in turn implies that the use of models and norms which can never be implemented or met by humans, due to limitations in their cognitive capacities or due to mismatches on a conceptual or motivational level, is only of limited interest.[3] Also, this stance does not give priority to any particular formal approach or modeling paradigm, but rather sees particular techniques and formalisms as means to an end with particular advantages and

[3] Clearly, this reading of rationality stands in fundamental conceptual disagreement with notions of AI and HLAI systems as "perfect reasoners" on a super-human level as envisioned, for instance, by Hutter (2005)'s AIXI model.

shortcomings, probably making necessary the incorporation of several different perspectives into an integrated or hybrid resulting framework.

As expanded upon in Besold and Kühnberger (2012), such a human-inspired notion of rationality can be founded on three conceptual cornerstones:

- **Subject centricity:** Rationality in a human context has to be considered as a subject-centered notion, demanding for the integration of subject-related properties and constraints.
- **Positive nature of the theory:** The main aspect of theories and models of rationality has to be their use and applicability as positive theories, and not as mere normative or postdictive-explanatory accounts.
- **Holistic nature of the approach:** A feasible theory of rationality does not have to be committed to one single formal modeling paradigm, but instead of being monolithic should pursue a holistic approach.

Taking all this together, this view opens the door for rationality to be seen (potentially) as a plethora of mechanisms competing, interacting and contributing to what can externally be observed as one single capacity which is strongly agent-specific in its precise form and manifestation. Still, the notion does not become trivialized as, due to the presumably considerable overlap in cognitive capacities and functions shared by all human reasoners, it can be expected that the basic characteristics should qualitatively be the same between individual subjects — which in turn should allow for the development of a general artificial model sharing these properties.

Besold (2013c) then grounds this notion in a strong relation to the context in which an agent faces a decision or choice situation and the (human) subject's ability to perform analogical reasoning. This contextualized analogy version of subject-centered rationality shares several basic properties with Rieskamp and Reimer (2007)'s notion of "ecological rationality", but is independent from the latter in its strong emphasis on the reasoner's abilities and their impact on the reasoning process rather than on the (almost automatic) guidance of the mental mechanisms by the environment.

Also, Besold (2013c) outlines how the HDTP computational analogy engine could be used as basis for a general architecture solving well-known rationality puzzles and, thus, for developing runnable models of the proposed notion of rationality: By going through a retrieval phase, in which analogical situations for a problem description and domain are retrieved, a mapping stage establishing correspondences between elements of the problem and the retrieved candidate situation(s), a transfer step in which solution-relevant knowledge is moved (and accordingly adapted) from the candidate situation(s) to the problem domain, and a final application phase in which the injected knowledge is then used for solving the problem in the corresponding domain. On a conceptual level, HDTP offers itself for an expansion in this direction as several of the internal processes and heuristics can directly be reused for guiding the just described process.

In summary, the work discussed in this section and the previous one contributes to understanding and developing HLAI and the role of analogy in HLAI systems in several ways: On the one hand, the introduced notion of subject-centered rationality allows for a more adequate understanding of and approach to rationality in the context of HLAI research and system engineering, possibly overcoming some of the hinderances and pitfalls which had been introduced by a strong fixation on particular formalisms and perfect, reasoner-independent rationality in the past. On the other hand, the importance for human-level intelligence of cognitive mechanisms, such as, for instance, analogy, and the resulting advantages of a cognitively-inspired perspective on HLAI have been further emphasized, and an architectural sketch for how HDTP as existing computational analogy framework could also be employed in the context of rationality has been given.

With respect to related efforts by other researchers, although the particular perspective on rationality and the approach to implementing it building on HDTP discussed above are new, an example for a related project on the computational side can be found in Petkov and Kokinov (2006)'s work on analogy-making, judgement and choice, or on the psychological side in the studies by Markman and Moreau (2001) on the role of analogy and analogy-based comparison in choice setups. Concerning the conceptual notion of subject-centered rationality, as already pointed out at the beginning of this section, some central elements are shared with the proposals by Simon (1956) and by Gilboa and Schmeidler (2001). Also, a substantial part of the underlying intuitions and ideas overlaps with the driving insights behind Cherniak (1986)'s proposal of "minimal rationality". Although the two accounts have been developed completely independently, the emphasis on the impact and ramifications of the boundedness of human cognition and reasoning as characteristic features of a meaningful account of rationality is common to both. Still, there are fundamental differences in the consequences drawn: Whilst Cherniak (1986) stays strongly rooted in the logical tradition of rationality, the proposal put forward by Besold (2013d), Besold and Kühnberger (2012), and Besold (2013c) does not commit to one particular formalism but argues in favor of an integrative pluralistic approach.

7. Complexity and Approximability as Guiding Forces in Human-Level Artificial Intelligence

Consideration of the computational properties, such as, for instance, the computational complexity, of a theory and corresponding models already played a central role in the notion of subject-centered rationality introduced in the previous chapter. Besold and Robere (2014) more generally advocate the use of similar formal methods, including complexity theory and approximation theory, as overall guidelines in the development of HLAI models and systems, with Besold (2014c) additionally pointing out the need for the development of a theory of structure-based approximation as necessary tool.[1] The suggestion is exemplified using the example of the parameterized complexity and approximability analysis of HDTP discussed in Chap. 3 and by a theoretical proposal for how to ground the notion of heuristics in cognitive systems and HLAI on solid formal foundations.

The fields of AI, HLAI, and cognitive systems research are characterized by a multitude of competing paradigms and approaches on all levels, from the high-level modeling of entire systems versus individual capacities, through the connectionist versus symbolist debate, down to questions of which particular logical language or which precise attractor model is superior over some other. Whilst this variety on the one hand should be seen as an advantage, as this far it is not clear which perspective or take on solving the (artificial) intelligence puzzle will succeed, on the other hand this often leads to parallel developments, incommensurable results and models, and a general lack of unifying principles and interfaces between the different streams.

Based on the recognition of this heterogeneity, Besold and Robere (2014) take inspiration from recent developments in theoretical cognitive science as basis of an attempt at establishing widely formalism- and design-independent standards and guiding forces for cognitive systems and HLAI research: As van Rooij (2008) points out, computational-level models of human cognitive capacities are most likely constrained by the computational complexity of the resulting theory — following the proposal given there, they have to be fixed-parameter tractable for parameters which can be assumed to be always small in practice. This has to be considered reasonable when entertaining a computational theory of mind as, among others, advocated by Pylyshyn (1980) and additionally accepting the

[1] See Part IV for a full version of Besold (2014c) and the final preprint of Besold and Robere (2014).

Church-Turing thesis (see, for instance, Turing (1969)), equating effective calculability with Turing computability: Against this background it seems highly unlikely that the mind could run some form of intractable theory in what we perceive as realtime.

Starting out from these and similar considerations, also additionally considering the so called "P-Cognition thesis" made known to an AI audience by Cooper (1990), Nebel (1996), and others, taking into account tractability and limitations on the admissible computational complexity of theories and models also in an HLAI and cognitive systems context seems recommendable. Following the example set by van Rooij (2008), Besold and Robere (2014) therefore introduce a "Tractable AGI thesis" which demands also for models of cognitive capacities in artificial intelligence and computational cognitive systems to be in FPT for parameters which can reasonably be assumed to be small in real-world scenarios.

The advantage of this form of high-level characterization of limiting factors for models and theories lies in its generality and independence from particular implementation details, by this overcoming many of the aforementioned divisions and barriers within HLAI and cognitive systems research and allowing for the introduction of general principles and requirements. The work by Robere and Besold (2012) and Besold and Robere (2013) moreover serves as a first worked out example of how a corresponding system analysis can look like and what kinds of results can, among others, be expected.

Continuing on a more theoretical level, Besold and Robere (2014) also provide a second example for how complexity theory and, additionally, approximation theory can be useful in establishing boundaries and guiding principles for work on computational models of cognition and intelligence. Building on the currently very popular practice of trying to use heuristics for solving complex problems in limited time, a formal characterization of the two main types of heuristics to be found in cognitive and HLAI systems is proposed: Reduction-based heuristics, trying to reduce a complex problem to a simpler, but solution-equivalent instance thereof, and approximation-based heuristics, trying to provide approximate solutions to the original problem. The former kind of heuristic is identified with the parameterized complexity concepts of kernelization and kernelizability, the latter is suggested to be taken as coinciding with the approximation-theoretical idea of constant-factor approximation algorithms APX.[2]

The kernelizability of a problem — which coincides with the problem's membership in FPT as discussed, for example, by Downey et al. (1997) — implies that there is a certain type of reduction of a problem instance to a commonly smaller or less-complex one (namely to an instance bounded in its size by the value of a certain computable function applied to the parameter of the original instance) within the same problem class which can then be solved instead of the original one. On the other hand, if a problem is not kernelizable no "downward reduction" of this kind exists. Setting this in relation to

[2] See Part II, Chap. B for a short overview of required basic notions from parameterized complexity theory and Chap. C for their counterparts from approximation theory.

the notion of reduction-based heuristics, in interpretation it can be said that this form of heuristic can only feasibly be applied to kernelizable problems as otherwise the existence of the hypothesized reduction cannot be guaranteed — which introduces a constraint on the class of problems that can be addressed with reduction-based heuristics in the first place, consequently equating this class with FPT.

Using polynomial-time APX approximability as standard for approximation-based heuristics, i.e., demanding for a problem to be approximable in polynomial time up to a certain constant factor of the optimal solution, can be motivated in a similar way: Only if it is possible to provide a factor for which the problem is polynomial-time approximable, and if this factor can be considered as meaningful in the overall context, it seems reasonable to assume that an approximation-based heuristic is in general suitable for solving the problem class.[3] Again, this delimits the class of problems which can reasonably be tackled with approximation-based heuristics from those for which this is not the case, this time using APX membership as separating criterion.

In a third step, taking an integrative perspective on heuristics by using the concept of fixed-parameter approximability FPA, both just discussed notions can be subsumed under what Besold and Robere (2014) call the "Fixed-Parameter Approximable AGI thesis". The latter, in the style of the earlier Tractable AGI thesis, demands for models of cognitive capacities in cognitive systems research and AI to be fixed-parameter approximable for input parameters that can reasonably be assumed to be small in practice — with the difference to the Tractable AGI thesis being that the FPA version also accommodates for approximation approaches. Conveniently, both notions previously used for characterizing the two distinct general types of heuristics, FPT and APX, in most cases imply membership in FPA.[4,5]

Whilst all the summarized proposals make use of "classical" complexity and approximation theory, Besold (2014c) additionally points out the need for another tool on the approximation side: Standard approaches to approximation are value-based in that they measure the degree of optimality of a solution provided by the approximation algorithm in terms of proximity of the value to the actual optimum for the optimization problem under consideration. Unfortunately, value similarity does not necessarily coincide with structure similarity — a proposed close-to-optimal solution in terms of value can be arbitrarily far from the real optimum in terms of structure. Whilst for many domains the value dimension is almost exclusively relevant, for cognitive systems and HLAI appli-

[3] Of course, also alternative classes to APX, such as PTAS, the class of problems for which there exists a polynomial-time approximation scheme, could be considered. Still, using APX seems to be the best choice in terms of balance between rigidity and flexibility of the notion, as well as in terms of its conceptual relation and theoretical and empirical fit to Simon (1956)'s notion of satisficing in human reasoning.

[4] For the FPT case, a slightly more specific form of FPT-membership, namely fixed-parameter tractability with witness as introduced in Cai and Chen (1997), is required.

[5] FPA also contains problems which are neither in FPT nor in APX, i.e., FPA is more general than the set-theoretical union between FPT and APX.

cations, such as, for instance, computational analogy-making as performed by HDTP, the structure of approximate solutions can play a crucial role. Unfortunately, although the need for such a tool has already been recognized, among others, by Hamilton et al. (2007), currently there is no theory or formal framework for performing structure-based approximation analysis, leaving a significant methodological gap that will have to be filled in the near future.

From a more general perspective, Besold and Robere (2014) continue the tradition started by work on the aforementioned P-Cognition thesis in AI in the late 1980s and early 1990s, additionally incorporating results of more recent developments in theoretical computer science, such as the tools and techniques offered by parameterized complexity theory and approximation theory. Concerning related work, whilst within (HL)AI and cognitive systems research currently only limited efforts are dedicated to further developing the formal basis and analytical methodological repertoire — with, for instance, Wareham et al. (2011)'s contribution on complexity issues in adaptive reactive architectures being a laudable exception —, theoretical cognitive science at the moment witnesses a growing number of contributions providing in-depth analysis of chances and limitations of computational-level approaches to modeling cognitive capacities. Examples for the latter line of research are, among others, Kwisthout et al. (2011) and Kwisthout and van Rooij (2012)'s results on approximate bayesian inference or Blokpoel et al. (2011)'s insights on the computational properties of certain communicative aspects. Concerning the question for structure-based approximation techniques introduced into the HLAI and cognitive systems debate by Besold (2014c), although in related fields of research already asked in the past, no solution seems to be in sight.

8. The Scientific Status of Human-Level Artificial Intelligence and the Testability Question

Having had a look at HLAI and the role cognitive capacities and computational models thereof can play in this line of research over the previous chapters, and before providing an overall concluding discussion in the next one, the focus now is shifted away from the question of how to implement and (re)create human-level intelligence in an artificial system, towards the question of what the scientific status of such an endeavor (and AI more generally) is and how its progress or success can be measured (if at all): Besold (2013b) addresses the general question for the scientific status of AI and HLAI research, Besold (2014a) provides a critical review of the chances and limitations of Bringsjord and Schimanski (2003)'s Psychometric AI proposal as quantitative evaluation standard for work in HLAI, and Besold (2013e) gives an alternate proposal in form of a revamped and refined version of Turing (1950)'s famous interactive testing paradigm.[1]

8.1. Human-Level Artificial Intelligence as a science

As pointed out by Besold (2013b), from a certain perspective AI, and even more HLAI, seems to stand out between the modern sciences for more than one reason: Neither is there an agreement upon what shall be AI's overall objective, nor is there a commonly accepted methodology for conducting research in AI and HLAI, nor is there consensus concerning the valuation of previous developments and of the actual status quo in AI as a story of success or perpetual failure.

These and related observations repeatedly caused philosophers of science and even some researchers from within AI to wonder about AI being a special type of science, or to even question (and occasionally finally deny) the status of AI and HLAI as a science. Recently, Cassimatis (2012) advocated that, when dealing with HLAI research, normal scientific standards and methods are often incidental and even antithetical to achieving human-level intelligence in an artificial system, and that a different approach would thus be required.

Cassimatis' line of arguments for why HLAI is not a normal science starts out with

[1] See Part IV for full versions of the referenced papers.

proclaiming a significant qualitative difference in the specificity and the level of ambition of the respective goals between research in HLAI and other sciences. From his point of view, the objectives of HLAI are "*more concrete and much more ambitious*" than the counterparts from other fields. Also, he claims that (HL)AI historically had not been conducted as a normal science as systems allegedly never witnessed experimental evaluation or formal proofs, which would be obligatory part of science or engineering research reports. Whilst the evaluation of the first claim seems to be a question of personal judgement only, the latter claim should at least be considered debatable: For instance, consider Winograd (1971)'s very early comparison of his SHRDLU system with other parsers and programs, or the numerous introspection reports that Newell and Simon collected and methodologically analyzed as basis for the development of their General Problem Solver Newell and Simon (1959). In doing so Newell and Simon implemented systematic observation, data collection, analysis and subsequent model building as classical methodological pattern from the natural sciences — and thus in their doing already back in the 1970s were compatible with the following, much more recent observation stated in Russell and Norvig (2009):

> "*In terms of methodology, AI has finally come firmly under the scientific method.*"

But the line of reasoning given by Cassimatis does not stop with claiming that HLAI is not a science in the normal sense, but continues with a much stronger and (if correct) even more troublesome observation. In his eyes, the application of scientific norms and methodological standards is detrimental to progress towards the goal of achieving HLAI. Two of Cassimatis' main quarrels consequently lie with formal or empirical demonstrations of correctness or optimality, and the connected computational requirements in terms of processing power and speed. From his perspective, the belief that importance should be assigned to showing the formal correctness of a theorem, or to empirically demonstrating a method's optimality with respect to a certain normative standard, goes against the nature of human intelligence. By reference to Simon (1956)'s notion of bounded rationality, Cassimatis tries to demonstrate that human rationality and reasoning fall far from optimality or formal correctness in all but a few cases, and that, thus, also HLAI should not use normative notions of correctness or optimality in judging and evaluating results.

As already pointed out in Chap. 6, it goes without doubt that human rationality is not optimal in any of the classical senses and is indeed subject to limitations arising from computational complexity or intractability. But this does not mean that no quantitative assessment is possible or meaningful per se: Whilst on the one hand a proposal for re-conceptualizing rationality in a more adequate way can be found in Chap. 6, and work towards a formal framework allowing to evaluate computational limitations and properties of HLAI frameworks is summarized in Chap. 7, on the other hand there is at least one approach within AI itself which claims to successfully apply quantitative measures to problems in HLAI research, i.e. research going by the name of Psychometric Artificial Intelligence (PAI).

8.2. What Psychometric AI can (not) do

Bringsjord and Schimanski (2003)'s PAI aims to apply the full battery of techniques from psychometrics to an HLAI context, setting its internal standard by declaring an agent as intelligent if and only if it does well in all established, validated tests of intelligence. This makes PAI a very quantitatively focused field of research with clear normative principles, but still seems to avoid the pitfalls Cassimatis (2012) meant to diagnose. As advocated by the proponents of PAI as methodology, the progress of an HLAI system towards achieving the overall goal of (re-)creating human intelligence is measured against actual human performance on what seems to be a commonly agreed upon standard means of assessing the relevant human mental capacities. By doing so, it is suggested that optimality is not demanded anymore with respect to a hypothetical idealized standard but with respect to achievable and reproducible testing scores of human subjects. Still, as shown by Besold (2014a), PAI falls short of its own overall expectations.

As introduced by Bringsjord and Schimanski (2003), the PAI approach can basically be divided into two streams with corresponding definitions of necessary and sufficient criteria for an artificial agent to be considered intelligent:

- **Naive Psychometric AI**: "*Some agent is intelligent if and only if it excels at all established, validated tests of intelligence.*"
- **General Psychometric AI**: "*Psychometric AI is the field devoted to building information-processing entities capable of at least solid performance on all established, validated tests of intelligence and mental ability, a class of tests that includes not just the rather restrictive IQ tests, but also tests of artistic and literary creativity, mechanical ability, and so on.*"

Whilst the shortcomings of the naive PAI proposal, such as a resulting overly narrow characterization of the full range of human cognitive capacities due to limitations of what is and can be evaluated by the named tests, become obvious quite quickly, the general PAI approach at first seems more promising. It seems to offer sufficient coverage in terms of addressed mental faculties as to assure that the passing HLAI would really match human-level standards, the standardized nature of the tests would allow the outcomes of different runs to be easily comparable, and the quantitative nature assessed on a continuous scale allows to not only measure success or failure but to also provide an indication of whether (and how quickly) the system is advancing towards meeting the test criteria or not.

Still, also general PAI is built on a possibly fatal misunderstanding: Namely Bringsjord and Schimanski (2003)'s misreading of psychometrics being the field which systematically measures, among others, intelligence in a direct manner. To the contrary, psychometric measures of intelligence are correlational measures testing the performance of traits commonly associated with what is interpreted as intelligence. In consequence, also general PAI does not train its systems on human-level intelligence but on correlated capacities without clear evidence that the resulting measure turns out to actually address intelligence in a reliable way. Similar to a patient suffering from savant syndrome, a system excelling on the scale of general PAI could still be limited to strong performance on particular cognitive tasks rather than exhibiting the desired general human-level intelligence (or beyond) — leaving even the best-willing behaviorist with unsurmountable doubts about whether the overall goal of HLAI has completely been achieved.

8.3. Turing revisited: The Sub-Turing Test(s)

In contrast to the psychometrics-based approach of the previous section, Besold (2013e) outlines a behavior-based test in the vein of Turing (1950)'s first proposal for a test for human-like capabilities of an artificial intelligence. The suggested Sub-Turing challenges for designing and testing artificially intelligent systems try to address human-level intelligence (or the recreation thereof in an artificial system) in terms of four interconnected sub tasks which, when taken together, are assumed to cover large parts of what would be needed to construct a serious competitor for solving HLAI: The first one addresses human language understanding, the second one human language production, the third one deals with human rationality and the fourth one refers to human operational or productive creativity. As criterion of whether the machine passes or fails the respective test a jury criterion putting the artificial system in competition with human subjects (i.e., closely resembling Turing's proposal) is introduced.

SubTuring I asks for the development of a system which should be capable of mapping natural language input onto expressions of a formal language and matching concepts in a lexical ontology, whilst SubTuring II demands the reversal of this process in that a set of formal language descriptions of situations, together with corresponding concepts in a lexical ontology, should be converted into human-like natural language output re-describing the situations. Out of the four sub tasks, the first two are probably the best studied this far: Among others disciplines, such as natural language processing, ontology engineering and (important parts of) data mining, are addressing closely related or sometimes even identical issues. Still, the challenges have not yet been solved and more targeted research in these areas will be needed.

SubTuring III, addressing human rationality, asks for a system to be capable of, provided with a set of situational descriptions and a problem or task description, solving or deciding a rationality problem/task. Similar to the argument made in Chap. 6, successfully solving SubTuring III would require a positive, computationally implementable and effectively runnable theory of human rationality instead of the mostly normative notions currently under study — here, cognitive modeling and, in general, cognitive science might give a very much needed helping hand.

SubTuring IV finally deals with human-level creativity: Provided with a description of a situation and an operational or productive creativity task or problem, the system would have to solve the problem/task in a human-like manner. Within AI, there is only little previous work trying to address creativity on a larger scale, and even within the dedicated field of computational creativity (of which an overview description has, for instance, been provided by Colton et al. (2009)) only very few projects or programs explicitly address these issues in a broader sense.

Taking all four SubTuring tasks together, and provided that some merit is granted to behavioral tests of the Turing type, the sketched proposal has at least one significant advantage over Turing (1950)'s original version: Whilst the latter provides only very little structure or hint at how to approach the problem, the SubTuring tasks introduce a division into four sub tasks all of which, from an engineering perspective, seem to be better defined than the mere domain- and faculty-unspecific human-likeness criterion as advocated by Turing. SubTuring I-IV contain an implicit commitment to what are (at least) necessary or (at best) sufficient compounds an artificial intelligence would have to contain for credibly addressing HLAI in its fullest generality — with the behavioral nature of the test rather being taken as an advantage than as a shortcoming given that there is no clear characterization of what intelligence as a cognitive capacity actually is or can be.

In summary, the work discussed in this chapter advocates for AI and HLAI to be taken serious in their scientific aspirations, and even more hints at the necessity of a scientific approach to the respective research questions. Another, more detailed case for AI and HLAI as (empirical) science has, for instance, been made by Simon (1995). Concerning the assessment and testability dimension, whilst the currently popular suggestion of using Psychometric AI as a foundation for making progress in HLAI measurable has been identified as not unproblematic and as in all likelihood falling short of its promised capabilities, the SubTuring tasks have instead been suggested as a refinement of the Turing Test. Concerning related efforts, especially with respect to alternatives or modifications to the original Turing idea, there is a significant body of work spearheaded, for example, by Harnad (1991) and his "Total Turing Test" (which demands in addition to the standard Turing criteria to also take perceptual abilities and the capability of physical object manipulation into account) or by McKinstry (1997)'s conceptually opposed "Minimum Intelligent Signal Test" (which only allows the binary responses yes/no or /true/false in order to to exclusively focus on the capacity for thought).

9. Conclusion

As already pointed out in the opening paragraphs of Chap. 1, a significant part of the reported work has to be considered exploratory in nature, testing the general suitability of paradigms across domains and the overall feasibility of approaches in different contexts by providing proof of concept applications and examples rather than taking one precise problem and developing one exact solution for that particular task until its possibilities on that one scenario have been exhausted. Consequently, whilst still tying into each other on different levels and in their interrelatedness suggesting answers to several foundational questions (such as “What could be guiding constraints for research and development in HLAI?” or “How can human high-level cognitive capacities, such as conceptual blending or decision-making, computationally be modeled in a unified manner?”), each of the different sub-projects also has an existence of its own and could be further developed in its own direction.

The work on conceptual blending reported on in Chap. 4 by now has been taken up in a coordinated European research project called “Concept Invention Theory (COINVENT)” aiming at developing a mathematically sound, computationally feasible and cognitively inspired formal model of concept invention. Concerning the studies on the use of computational analogy engines for modeling analogy-based methods and tools from education and teaching discussed in Chap. 5, a logical next step would be to compare the results from the computational model to quantitative psychological data and to better tune the models to match the experimental data — which hopefully would allow to subsequently use the analogy engine also for predictions about the success or failure of new educational analogies or for the improvement of existing ones. The efforts on the use of analogy (and computational models thereof) in explaining, modeling and (re)implementing human decision-making, judgment and in general behavior, described in Chap. 6, as a next step would also profit from an actual implementation of the proposed theory and architecture and an application of the outcome to experimental data from decision studies and choice experiments. With respect to the theoretical developments sketched in Chap. 7, two directions offer themselves for follow-up investigations: On the one hand, thorough analysis of existing cognitive architectures and candidate systems for becoming an HLAI is needed in order to provide further examples of the chances and usefulness the proposed forms of formal examination offer (and also in order to separate more promising from less auspicious projects and possibly refocus efforts within the field), whilst on the other hand the existing formal machinery should be further developed either by introducing more already existing but widely ignored tools and techniques to a wider audience, or by actively working on new approaches, such as the mentioned structure-based approximation methods. Finally, returning to Chap. 8, AI and HLAI as every other scientific endeavor needs means of evaluating its progress and deciding over success or failure, therefore further proposals for meaningful and feasible ways of testing the state of development of a cognitive system or an HLAI project are needed.

In summary, the main merit of work reported in this thesis is twofold: On the one hand, using analogy as starting point, examples are given as inductive evidence for how a cognitively-inspired approach to questions in HLAI can be fruitful by and within itself. On the other hand, (some of) the advantages of such a perspective also with respect to overcoming certain intrinsic problems currently characterizing HLAI research in its entirety — such as the domain dependence and lack of generalizability of specialized solution methods or the reliance on raw computational power instead of versatile and efficient solution mechanisms — are exposed. In terms of individual outcomes, an analogy-based proposal for theory blending as special form of conceptual blending is introduced and exemplified, the usefulness of computational analogy frameworks for understanding learning and education is shown and a corresponding research program is suggested, a subject-centered notion of rationality and a sketch for how the resulting theory could computationally be modeled using an analogy framework is discussed, computational complexity and approximability considerations are introduced as guiding principles for work in HLAI, and the scientific status of HLAI, as well as two possible tests for assessing progress in HLAI, are addressed.

I want to close with two quotes, one by Papadimitriou and Yannakakis (1991) and one by Turing (1950):

> "*Once more, we have decreased the number of open questions in the field — without, alas, increasing much the number of answers!*" (C. H. Papadimitriou and M. Yannakis)
>
> "*We can only see a short distance ahead, but we can see plenty there that needs to be done.*" (A. Turing)

Part II.

Appendix

A. Overview of Basic Mechanisms of Heuristic-Driven Theory Projection

According to Schwering et al. (2009a) and others, the HDTP framework has been conceived as a mathematically sound theoretical framework and implemented engine for computational analogy-making. As further explained by Guhe et al. (2011), HDTP more precisely has been created for computing analogical relations and inferences for domains which are given in form of many-sorted FOL representations: Source and target of the analogy-making process are defined in terms of axiomatizations, i.e., given by a finite set of formulae. From there, HDTP tries to align pairs of formulae from the two domains by means of anti-unification. Anti-unification, as introduced by Plotkin (1970) and Plotkin (1971), is the dual to the more prominent unification problem. Basically, anti-unification tries to solve the problem of generalizing terms in a meaningful way, yielding for each term an anti-instance, in which distinct sub-terms have been replaced by variables (which in turn would allow for a retrieval of the original terms by a substitution of the variables by appropriate sub-terms).

The goal of anti-unification is to find a most specific anti-unifier, i.e., the least general generalization of the involved terms.[1] HDTP extends first-order anti-unification to a restricted form of higher-order anti-unification, as mere first-order structures must be considered as too weak for the purpose of analogy-making: Structural commonalities can be embedded in different contexts, and therefore would not be accessible by first-order anti-unification only.

In Krumnack et al. (2007)'s conceptualization of restricted higher-order anti-unification a new notion of substitution is introduced in order to restrain generalizations from becoming arbitrarily complex. Classical first-order terms are extended by the introduction of variables which may take arguments (where original first-order variables correspond to variables with arity 0), making a term either a first-order or a higher-order term. Subsequently, anti-unification can be applied analogously to the original first-order case, yielding a generalization subsuming the specific terms.

As already indicated by the name, the class of substitutions which are applicable in HDTP is restricted to (compositions of) the following four cases: renamings, fixations, argument insertions, and permutations.

[1] Plotkin (1970) has shown that for a proper definition of generalization, for a given pair of terms there always is a first-order generalization, and that there is exactly one least general first-order generalization (up to renaming of variables).

Definition. Substitutions in Restricted Higher-Order Anti-Unification
Let $\mathcal{V} = \{x_1 : s_1, x_2 : s_2, \ldots\}$ be an infinite set of sorted variables, where the sorts are chosen from a set of sorts *Sort*. Associated with each variable $x_i : s_i$ is an arity, analogous to the standard arity of function symbols. For any $i \geq 0$, we let $\mathcal{V}_i$ be the variables of arity i.

1. A renaming $\rho(F, F')$ replaces a variable $F \in \mathcal{V}_n$ with another variable $F' \in \mathcal{V}_n$:
$F(t_1, \ldots, t_n) \xrightarrow{\rho(F,F')} F'(t_1, \ldots, t_n)$.

2. A fixation $\phi(F, f)$ replaces a variable $F \in \mathcal{V}_n$ with a function symbol $f \in \mathcal{C}_n$:
$F(t_1, \ldots, t_n) \xrightarrow{\phi(F,f)} f(t_1, \ldots, t_n)$.

3. An argument insertion $\iota(F, F', V, i)$ is defined as follows, where $F \in \mathcal{V}_n, F' \in \mathcal{V}_{n-k+1}, V \in \mathcal{V}_k, i \in [n]$:
$F(t_1, \ldots, t_n) \xrightarrow{\iota(F,F',V,i)} F'(t_1, \ldots, t_{i-1}, V(t_i, \ldots, t_{i+k-1}), t_{i+k}, \ldots, t_n)$.
It “wraps” k of the subterms in a term using a k-ary variable, or can be used to insert a 0-ary variable.

4. A permutation $\pi(F, \tau)$ rearranges the arguments of a term, with $F \in \mathcal{V}_n$, $\tau : [n] \to [n]$ a bijection:
$F(t_1, \ldots, t_n) \xrightarrow{\pi(F,\tau)} F(t_{\tau(1)}, \ldots, t_{\tau(n)})$.

A restricted substitution is a substitution which results from the composition of any sequence of unit substitutions.

Krumnack et al. (2007) show that this new form of (higher-order) substitution is a real extension of the first-order case, which has proven to be capable of detecting structural commonalities not accessible to first-order anti-unification. In some form of trade-off, the least general generalization unfortunately loses its uniqueness. Therefore, HDTP ranks generalizations according to a complexity order on the complexity of generalization (based on a complexity measure for substitutions), and finally chooses the least complex generalizations as preferred ones. From a practical point of view, it is also necessary to anti-unify not only terms, but formulae: HDTP extends the notion of generalization also to formulae by basically treating formulae in clause form and terms alike (as positive literals are structurally equal to function expressions, and complex clauses in normal form may be treated component wise).

Furthermore, analogies in general not only rely on an isolated pair of formulae from source and target, but on two sets of formulae. Here, a heuristic is applied when iteratively selecting pairs of formulae to be generalized: Coherent mappings outmatch incoherent ones, i.e., mappings in which substitutions can be reused are preferred over isolated substitutions, as they are assumed to be better suited to induce the analogical relation. Once obtained, the generalized theory and the substitutions specify the analogical relation, and formulae of the source for which no correspondence in the target domain can be found may, by means of the already established substitutions, be transferred to the target, constituting a process of analogical transfer between the domains.

For a more in-depth introduction to the framework and the actual system, see, for instance, Schmidt et al. (2014).

B. Overview of Basic Definitions From Parameterized Complexity Theory

Definition. Parameterized Decision Problem
An instance of a parameterized decision problem $\mathcal{P}$ is a tuple (x, κ), where $x \in \{0,1\}^*$ is a string describing the problem and $\kappa \in \mathbb{Z}$, which is called the parameter of the problem (codifying other aspects of the problem besides n).

Similar to the use of the complexity class P in standard complexity theory, tractability in parameterized complexity theory is defined using the class FPT:

Definition. Fixed-Parameter Tractability FPT
A parameterized decision problem $\mathcal{P}$ is fixed parameter tractable, written $\mathcal{P} \in \mathsf{FPT}$, if it is solvable in time bounded by $f(\kappa) \cdot |x|^{O(1)}$, where $f(\kappa)$ is some computable function of the parameters and $|x|^{O(1)}$ denotes a polynomial of the length of the input.

Or, equivalently:

Definition. Fixed-Parameter Tractability FPT
A parameterized decision problem $\mathcal{P}$ is in FPT if $\mathcal{P}$ admits an $O(f(\kappa)n^c)$ algorithm, where n is the input size, κ is a parameter of the input, c is an independent constant, and f is some computable function.

Similar to standard complexity theory, parameterized complexity theory also applies reductions as one of its main tools:

Definition. Parameterized Reduction:
Given two parameterized problems $\mathcal{P}$, $\mathcal{Q}$, a parameterized reduction is a function ϕ from $\mathcal{P}$ to $\mathcal{Q}$ such that the following holds, for an instance $(x, \kappa) \in \mathcal{P}$:
1.) $\phi(x)$ is computable in time $f(\kappa) \cdot |x|^{O(1)}$, where f is a computable function of the parameter.
2.) $x \in P$ if and only if $\phi(x) \in Q$.
3.) If κ' is the parameter of $\phi(x)$, then $\kappa' = g(\kappa)$ for some function g.

The parameterized complexity hierarchy is established using different versions of Weighted Circuit Satisfiability as reference problem for the respective classes:

Let the weft of a boolean circuit (containing only NOT gates, small AND and OR gates of fan-in ≤ 2 and large AND or OR gates of arbitrary finite fan-in) be the maximum number of large gates on any path from an input to the output, and let the depth be the maximum number of all gates on a path. Let $\mathsf{C[t,d]}$ be the set of all circuits of weft at most t and depth at most d. Finally, define the (Hamming) weight of an assignment of truth values to the input variables of the circuit as the number of variables set to 1.

Problem. Weighted Circuit Satisfiability$[t, d]$
Input: A circuit C of depth d and weft t, a natural $k \in \mathbb{N}$.
Problem: Is there a satisfying assignment to C with weight k?

Definition. W Hierarchy
We say a parameterized problem $\mathcal{P}$ is in $\mathsf{W}[i]$ if it is reducible by a parameterized reduction to Weighted Circuit Satisfiability$[i, d]$ for some constant d, and is $\mathsf{W}[i]$-hard if every problem in $W[i]$ is reducible to $\mathcal{P}$ under a parameterized reduction.

Knowing that $\mathsf{FPT} = \mathsf{W}[0]$, the assumption that $\mathsf{FPT} \neq \mathsf{W}[1]$ can be seen as analogous to the assumption that $\mathsf{P} \neq \mathsf{NP}$ in standard complexity theory.
It is conjectured that $\mathsf{W}[i] \subsetneq W[j]$ for any $i < j$.
The FPT membership of a problem entails the reducibility of each instance to a problem kernel (and vice versa):

Definition. Kernelization
Let $\mathcal{P}$ be a parameterized problem. A kernelization of $\mathcal{P}$ is an algorithm which takes an instance x of $\mathcal{P}$ with parameter κ and maps it in polynomial time to an instance y such that $x \in \mathcal{P}$ if and only if $y \in \mathcal{P}$, and the size of y is bounded by $f(\kappa)$ (where f is some computable function).

Definition. Kernelizability
A problem $\mathcal{P}$ is in FPT if and only if it is kernelizable.

For a more in-depth introduction to parameterized complexity, see, for instance, Downey and Fellows (1999) or Flum and Grohe (2006).

C. Overview of Basic Definitions From Approximation Theory

Approximation theory, similar to complexity theory, also works with problem classes and a corresponding hierarchy. One of the most important classes towards the lower end of the hierarchy is PTAS, the class of problems admitting polynomial-time approximation schemes:

Definition. Polynomial-Time Approximability PTAS
An optimization problem P is in PTAS if for each parameter $\epsilon > 0$ there is an algorithm which takes an instance of P of size n together with ϵ and, in time polynomial in n, produces a solution that is within a factor $1+\epsilon$ of being optimal (or $1-\epsilon$ for maximization problems).

Located one major step higher in the class hierarchy is APX, the class of constant-factor approximable problems:

Definition. Constant-Factor Approximability APX
An optimization problem P is in APX if P admits a constant-factor approximation algorithm, i.e., there is a constant factor $\epsilon > 0$ and an algorithm which takes an instance of P of size n and, in time polynomial in n, produces a solution that is within a factor $1+\epsilon$ of being optimal (or $1-\epsilon$ for maximization problems).

Supposing $\mathsf{P} \neq \mathsf{NP}$, it holds that $\mathsf{PTAS} \subsetneq \mathsf{APX}$.
As shown by Cai and Huang (2006), APX membership also implies membership in the more general class of fixed-parameter approximable problems FPA:

Definition. Fixed-Parameter Approximabiliy FPA
The fixed-parameter version P of a minimization problem is in FPA if — for a recursive function f, a constant κ, and some fixed recursive function g — there exists an algorithm such that for any given problem instance I with parameter κ, and question $OPT(I) \leq \kappa$, the algorithm which runs in $O(f(\kappa)n^c)$ (where $n = |I|$) either outputs “no” or produces a solution of cost at most $g(\kappa)$.

For a more in-depth introduction to approximation theory, see, for instance, Vazirani (2010) or Marx (2008) for a more specific overview of work at the intersection between parameterized complexity and approximation theory.

Part III.

Bibliography

Bibliography

Abdel-Fattah, A. M., Besold, T. R., and Kühnberger, K.-U. (2012). Creativity, cognitive mechanisms, and logic. In Bach, J., Goertzel, B., and Ikle, M., editors, *Artificial General Intelligence*, volume 7716 of *Lecture Notes in Computer Science*, pages 1–10. Springer.

Abdel-Fattah, A. M. H. and Krumnack, U. (2013). Creating Analogy-Based Interpretations of Blended Noun Concepts. In *Proceedings of the AAAI Spring 2013 Symposium on Creativity and (Early) Cognitive Development*, AAAI Press Technical Reports.

Abdel-Fattah, A. M. H., Krumnack, U., and Kühnberger, K.-U. (2013). Utilizing Cognitive Mechanisms in the Analysis of Counterfactual Conditionals by AGI Systems. In Kühnberger, K.-U., Rudolph, S., and Wang, P., editors, *Artificial General Intelligence*, volume 7999 of *Lecture Notes in Computer Science*, pages 1–10. Springer.

Akgul, E. (2006). Teaching Science In An Inquity-Based Learning Environment: What It Means For Pre-Service Elementary Science Teachers. *Eurasia Journal of Mathematics, Science and Technology Education*, 2(1):71–81.

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, 149(1):91 – 130.

Argand, J.-R. (1813). Philosophie mathématique. Essai sur une manière de représenter les quantités imaginaires, dans les constructions géométriques, journal = Annales de Mathématiques pures et appliquées. 4:133–146.

Arnold, M. and Millar, R. (1996). Exploring the use of analogy in the teaching of heat, temperature and thermal equilibrium. In Welford, G., Osborne, J., and Scott, P., editors, *Research in Science Education in Europe: Current Issues and Themes*. Farmer Press, London.

Besold, T. R. (2011). Computational Models of Analogy-Making. An Overview Analysis of Computational Approaches to Analogical Reasoning. Technical Report X-2011-03, FNWI/FGw: Institute for Logic, Language and Computation (ILLC), University of Amsterdam.

Besold, T. R. (2013a). Analogy Engines in Classroom Teaching: Modeling the String Circuit Analogy. In *Proceedings of the AAAI Spring 2013 Symposium on Creativity and (Early) Cognitive Development*, AAAI Press Technical Reports.

Besold, T. R. (2013b). Human-Level Artificial Intelligence Must Be a Science. In Kühnberger, K.-U., Rudolph, S., and Wang, P., editors, *Artificial General Intelligence - 6th International Conference, AGI 2013, Proceedings*, volume 7999 of *Lecture Notes in Computer Science*, pages 174–177. Springer.

Besold, T. R. (2013c). Rationality in context: An analogical perspective. In Brezillon, P., Blackburn, P., and Dapoigny, R., editors, *Modeling and Using Context*, volume 8175 of *Lecture Notes in Computer Science*, pages 129–142. Springer.

Besold, T. R. (2013d). Rationality in|for|through AI. In Kelemen, J., Romportl, J., and Zackova, E., editors, *Beyond Artificial Intelligence*, volume 4 of *Topics in Intelligent Engineering and Informatics*, pages 49–62. Springer.

Besold, T. R. (2013e). Turing Revisited: A Cognitively-Inspired Decomposition. In Müller, V. C., editor, *Philosophy and Theory of Artificial Intelligence*, pages 121–132. Springer.

Besold, T. R. (2014a). A Note on Chances and Limitations of Psychometric AI. In Lutz, C. and Thielscher, M., editors, *KI 2014: Advances in Artificial Intelligence, Proceedings of the 37th Annual German Conference on AI*, volume 8736 of *Lecture Notes in Computer Science*. Springer.

Besold, T. R. (2014c). Towards Formally Well-Founded Heuristics in Cognitive AI Systems. *Cognitive Processing*, 15(1 Supplement).

Besold, T. R. (forthcoming, 2014b). Sensorimotor Analogies in Learning Abstract Skills and Knowledge: Modeling Analogy-Supported Education in Mathematics and Physics. In *Proceedings of the AAAI Fall 2014 Symposium on Modeling Changing Perspectives: Reconceptualizing Sensorimotor Experiences*, AAAI Press Technical Reports.

Besold, T. R. and Kühnberger, K.-U. (2012). E Pluribus Multa In Unum: The Rationality Multiverse. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Besold, T. R. and Kühnberger, K.-U. (2014). Applying AI for Modeling and Understanding Analogy-Based Classroom Teaching Tools & Techniques. In Lutz, C. and Thielscher, M., editors, *KI 2014: Advances in Artificial Intelligence, Proceedings of the 37th Annual German Conference on AI*, volume 8736 of *Lecture Notes in Computer Science*. Springer.

Besold, T. R., Pease, A., and Schmidt, M. (2013). Analogy and Arithmetics: An HDTP-Based Model of the Calculation Circular Staircase. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Besold, T. R. and Robere, R. (2013). When Almost Is Not Even Close: Remarks on the Approximability of HDTP. In Kühnberger, K.-U., Rudolph, S., and Wang, P., editors, *Artificial General Intelligence - 6th International Conference, AGI 2013, Proceedings*, volume 7999 of *Lecture Notes in Computer Science*, pages 11–20. Springer.

Besold, T. R. and Robere, R. (forthcoming, 2014). When Thinking Never Comes to a Halt: Using Formal Methods in Making Sure Your AI Gets the Job Done Good Enough. In Müller, V. C., editor, *Fundamental Issues of Artificial Intelligence*, Synthese Library. Springer.

Blokpoel, M., Kwisthout, J., Wareham, T., Haselager, P., Toni, I., and van Rooij, I. (2011). The computational costs of recipient design and intention recognition in communication. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 465–470. Cognitive Science Society.

Boden, M. A. (2003). *The Creative Mind: Myths and Mechanisms.* Routledge.

Bringsjord, S. and Schimanski, B. (2003). What is Artificial Intelligence? Psychometric AI as an Answer. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03).* Morgan Kaufmann.

Broadie, S. and Rowe, C., editors (2002). *Aristotle Nicomachean Ethics: Translation, Introduction, and Commentary.* Oxford University Press.

Cai, L. and Chen, J. (1997). On fixed-parameter tractability and approximability of {NP} optimization problems. *Journal of Computer and System Sciences*, 54(3):465 – 474.

Cai, L. and Huang, X. (2006). Fixed-parameter approximation: Conceptual framework and approximability results. In Bodlaender, H. and Langston, M., editors, *Parameterized and Exact Computation*, volume 4169 of *Lecture Notes in Computer Science*, pages 96–108. Springer.

Cassimatis, N. I. (2012). Human-Level Artificial Intelligence Must Be an Extraordinary Science. *Advances in Cognitive Systems*, 1:37–45.

Cherniak, C. (1986). *Minimal Rationality.* MIT Press.

Chrisley, R. (2003). Embodied artificial intelligence. *Artificial Intelligence*, 149(1):131–150.

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *Journal of Research in Science Teaching*, 30:1241–1257.

Colton, S., Lopez de Mantaras, R., and Stock, O. (2009). Computational Creativity: Coming of Age. *AI Magazine*, 30(3):11–14.

Cooper, G. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405.

Downey, R. G. and Fellows, M. R. (1999). *Parameterized Complexity.* Springer.

Downey, R. G., Fellows, M. R., and Stege, U. (1997). Parameterized complexity: A framework for systematically confronting computational intractability. In *Contemporary Trends in Discrete Mathematics: From DIMACS and DIMATIA to the Future.* AMS.

Doyle, J. (1992). Rationality and its role in reasoning. *Computational Intelligence*, 8(2):376–409.

Duit, R. (1991). The role of analogies and metaphors in learning science. *Science Education*, 75(6):649–672.

Evans, J. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, 128:978–996.

Evans, T. G. (1964). A heuristic program to solve geometric-analogy problems. In *Proceedings of the April 21-23, 1964, Spring Joint Computer conference*, AFIPS '64 (Spring), pages 327–338, New York, NY, USA. ACM.

Falkenhainer, B., Forbus, K., and Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1 – 63.

Fauconnier, G. and Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, 22(2):133–187.

Fauconnier, G. and Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities.* Basic Books, New York.

Fauconnier, G. and Turner, M. (2008). Rethinking metaphor. In Gibbs, R., editor, *Cambridge Handbook of Metaphor and Thought*, pages 53–66. Cambridge University Press, New York.

Flum, J. and Grohe, M. (2006). *Parameterized Complexity Theory.* Springer.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Gentner, D., Bowdle, B., Wolff, P., and Boronat, C. (2001). Metaphor Is Like Analogy. In Gentner, D., Holyoak, K., and Kokinov, B., editors, *The Analogical Mind: Perspectives from Cognitive Science*, pages 199–253. MIT Press.

Gentner, D., Falkenhainer, B., and Skorstad, J. (1988). Viewing metaphor as analogy. In Helman, D., editor, *Analogical Reasoning*, volume 197 of *Synthese Library*, pages 171–177. Springer Netherlands.

Gentner, D. and Forbus, K. (1991). MAC/FAC: A Model of Similarity-based Retrieval. *Cognitive Science*, 19:141–205.

Gentner, D. and Markman, A. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52(1):4 –56.

Gentner, D. and Smith, L. A. (2013). Analogical learning and reasoning. In Reisberg, D., editor, *The Oxford Handbook of Cognitive Psychology*, pages 668–681. Oxford University Press, New York, NY, USA.

Gigerenzer, G., Hertwig, R., and Pachur, T., editors (2011). *Heuristics: The Foundation of Adaptive Behavior.* Oxford University Press.

Gilboa, I. and Schmeidler, D. (2001). *A Theory of Case-Based Decisions.* Cambridge University Press.

Goguen, J. (2006). Mathematical models of cognitive space and time. In Andler, D., Ogawa, Y., Okada, M., and Watanabe, S., editors, *Reasoning and Cognition; Proceedings of the Interdisciplinary Conference on Reasoning and Cognition*, pages 125–128.

Goguen, J. A. and Harrell, D. F. (2010). Style: A computational and conceptual blending-based approach. In Argamon, S., Burns, K., and Dubnov, S., editors, *The Structure of Style*, pages 291–316. Springer.

Goswami, U. (2001). Analogical reasoning in children. In Gentner, D., Holyoak, K., and Kokinov, B., editors, *The analogical mind: Perspectives from cognitive science*, pages 437–470. MIT Press, Cambridge, MA.

Griffiths, T., Kemp, C., and Tenenbaum, J. (2008). Bayesian models of cognition. In Sun, R., editor, *The Cambridge Handbook of Computational Cognitive Modeling.* Cambridge University Press.

Guerra-Ramos, M. (2011). Analogies as Tools for Meaning Making in Elementary Science Education: How Do They Work in Classroom Settings? *Eurasia Journal of Mathematics, Science and Technology Education*, 7(1):29–39.

Guhe, M., Pease, A., Smaill, A., Martinez, M., Schmidt, M., Gust, H., Kühnberger, K.-U., and Krumnack, U. (2011). A computational account of conceptual blending in basic mathematics. *Journal of Cognitive Systems Research*, 12(3):249–265.

Guhe, M., Pease, A., Smaill, A., Schmidt, M., Gust, H., Kühnberger, K.-U., and Krumnack, U. (2010). Mathematical reasoning with higher-order anti-unification. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1992–1997.

Gust, H., Kühnberger, K.-U., and Schmid, U. (2006). Metaphors and Heuristic–Driven Theory Projection (HDTP). *Theoretical Computer Science*, 354:98–117.

Hamilton, M., Müller, M., van Rooij, I., and Wareham, T. (2007). Approximating Solution Structure. In *Dagstuhl Seminar Proceedings Nr. 07281*. IBFI, Schloss Dagstuhl.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1:45–54.

Hofstadter, D. (2001). Epilogue: Analogy as the core of cognition. In Gentner, D., Holyoak, K., and Kokinov, B., editors, *The Analogical Mind: Perspectives from Cognitive Science*, pages 499–538, Cambridge, MA. MIT Press.

Hofstadter, D. and Mitchell, M. (1994). The copycat project: a model of mental fluidity and analogy-making. In Holyoak, K. and Barnden, J., editors, *Advances in Connectionist and Neural Computation Theory*, volume 2: Analogical Connections, pages 31–112, New York, NY, USA. Ablex.

Holyoak, K. and Thagard, P. (1995). *Mental leaps: Analogy in creative thought.* MIT Press, Cambridge, MA.

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability.* Springer.

Krumnack, U., Schwering, A., Gust, H., and Kühnberger, K. (2007). Restricted Higher-Order Anti-Unification for Analogy Making. In *Twentieth Australian Joint Conference on Artificial Intelligence.* Springer.

Kühnberger, K.-U., Geibel, P., Gust, H., Krumnack, U., Ovchinnikova, E., Schwering, A., and Wandmacher, T. (2008). Learning from inconsistencies in an integrated cognitive architecture. In Wang, P., Goertzel, B., and Franklin, S., editors, *Artificial General Intelligence 2008, Proceedings of the First AGI Conference*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pages 212–223. IOS Press.

Kühnberger, K.-U., Wandmacher, T., Schwering, A., Ovchinnikova, E., Krumnack, U., Gust, H., and Geibel, P. (2007). I-cog: A computational framework for integrated cognition of higher cognitive abilities. In Gelbukh, A. and Kuri Morales, A. F., editors, *MICAI 2007: Advances in Artificial Intelligence*, volume 4827 of *Lecture Notes in Computer Science*, pages 203–214. Springer.

Kwisthout, J. and van Rooij, I. (2012). Bridging the gap between theory and practice of approximate bayesian inference. In *Proceedings of the 11th International Conference on Cognitive Modeling*, pages 199–204.

Kwisthout, J., Wareham, T., and van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, 35(5):779–784.

Markman, A. and Moreau, C. (2001). Analogy and analogical comparison in choice. In Gentner, D., Holyoak, K., and Kokinov, B., editors, *The Analogical Mind: Perspectives from Cognitive Science*, pages 363–399. MIT Press.

Marr, D. (1982). *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information.* W. H. Freeman and Company.

Martinez, M., Besold, T. R., Abdel-Fattah, A., Gust, H., Schmidt, M., Krumnack, U., and Kühnberger, K.-U. (2012). Theory blending as a framework for creativity in systems for general intelligence. In Wang, P. and Goertzel, B., editors, *Theoretical Foundations of Artificial General Intelligence*, volume 4 of *Atlantis Thinking Machines*, pages 219–239. Atlantis Press.

Martinez, M., Besold, T. R., Abdel-Fattah, A., Kühnberger, K.-U., Gust, H., Schmidt, M., and Krumnack, U. (2011). Towards a domain-independent computational framework for theory blending. In *Proceedings of the AAAI Fall 2011 Symposium on Advances in Cognitive Systems*, AAAI Press Technical Reports.

Marx, D. (2008). Parameterized complexity and approximation algorithms. *The Computer Journal*, 51(1):60–78.

McCarthy, J. (2007). From Here to Human-Level AI. *Artificial Intelligence*, 171:1174–1182.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, 27:12–14.

McKinstry, C. (1997). Minimum intelligent signal test: An alternative turing test. *Canadian Artificial Intelligence*, 41.

Moore, R. (1995). *Logic and Representation.* Cambridge University Press, Cambridge, England.

Nebel, B. (1996). Artificial intelligence: A computational perspective. In Brewka, G., editor, *Principles of knowledge representation*, pages 237–266. CSLI Publications.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4:135–183.

Newell, A. and Simon, H. (1959). Report on a general problem-solving program. In *Proceedings of the International Conference on Information Processing*.

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence*. Cambridge University Press, New York, NY, USA.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental psychology: Learning, Memory, and Cognition*, 14:510–520.

Osborne, M. and Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press.

Papadimitriou, C. H. and Yannakakis, M. (1991). Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences*, 43(3):425 – 440.

Pereira, F. C. (2007). *Creativity and AI: A Conceptual Blending Approach*. Applications of Cognitive Linguistics (ACL). Mouton de Gruyter, Berlin.

Petkov, G. and Kokinov, B. (2006). JUDGEMAP - integration of analogy-making, judgement, and choice. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1950–1955. Cognitive Science Society.

Plotkin, G. D. (1970). A note on inductive generalization. *Machine Intelligence*, 5:153–163.

Plotkin, G. D. (1971). A further note on inductive generalization. *Machine Intelligence*, 6:101–124.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundation of cognitive science. *Behavioral and Brain Sciences*, 3(1):111–132.

Reitman, W. R., Grove, R. B., and Shoup, R. G. (1964). Argus: An information-processing model of thinking. *Behavioral Science*, 9(3):270–281.

Rieskamp, J. and Reimer, T. (2007). Ecological rationality. In Baumeister, R. and Vohs, K., editors, *Encyclopedia of Social Psychology*, pages 273–275. Sage.

Robere, R. and Besold, T. R. (2012). Complex Analogies: Remarks on the Complexity of HDTP. In *Twentyfifth Australasian Joint Conference on Artificial Intelligence*, volume 7691 of *Lecture Notes in Computer Science*, pages 530–542. Springer.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13:629–639.

Rumelhart, D. E., Hinton, G. E., and McClelland, J. L. (1986). A general framework for parallel distributed processing. In Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pages 45–76. MIT Press, Cambridge, MA, USA.

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence*, 94(1–2):57–77. Economic Principles of Multi-Agent Systems.

Russell, S. J. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ.

Schmidt, M., Gust, H., Kühnberger, K.-U., and Krumnack, U. (2011). Refinements of Restricted Higher-Order Anti-Unification for Heuristic-Driven Theory Projection. In *KI 2011: Advances in Artificial Intelligence, Proceedings of the 34th Annual German Conference on Artificial Intelligence*, volume 7006 of *Lecture Notes in Computer Science*. Springer.

Schmidt, M., Krumnack, U., Gust, H., and Kühnberger, K.-U. (2014). Heuristic-driven theory projection: An overview. In Prade, H. and Richard, G., editors, *Computational Approaches to Analogical Reasoning: Current Trends*, pages 163–194. Springer.

Schnalle, K. and Schwank, I. (2006). Das Zahlen-Hochhaus [ZH]: Multiplikative Zusammenhänge im Hunderterraum. In *Beiträge zum Mathematikunterricht*. Franzbecker.

Schwank, I., Aring, A., and Blocksdorf, K. (2005). Betreten erwünscht - die Rechenwendeltreppe. In *Beiträge zum Mathematikunterricht*. Franzbecker, Hildesheim.

Schwering, A., Gust, H., and Kühnberger, K.-U. (2009a). Solving Geometric Analogies with the Analogy Model HDTP. In Taatgen, N. A. and van Rijn, H., editors, *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, pages 1780–1785. Cognitive Science Society.

Schwering, A., Krumnack, U., Kühnberger, K.-U., and Gust, H. (2009b). Syntactic principles of Heuristic-Driven Theory Projection. *Journal of Cognitive Systems Research*, 10(3):251–269.

Schwering, A., Kühnberger, K.-U., and Kokinov, B. (2009c). Analogies: Integrating multiple cognitive abilities - guest editorial. *Journal of Cognitive Systems Research*, 10(3).

Shelley, C. (2003). *Multiple Analogies in Science and Philosophy*. John Benjamins Publishing, Amsterdam; Philadelphia.

Siegler, R. (1989). Mechanisms of Cognitive Development. *Annual Review of Psychology*, 40:353–379.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63:129–138.

Simon, H. A. (1995). Explaining the Ineffable: AI on the Topics of Intuition, Insight and Inspiration. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95) - Volume 1*, pages 939–948. Morgan Kaufmann Publishers Inc.

Smolensky, P. (1987). Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, 1(2):95–109.

Stedall, J. (2008). *Mathematics emerging: A Sourcebook 1540–1900.* Oxford University Press, Oxford.

Thagard, P., Cohen, D., and Holyoak, K. (1989). Chemical Analogies: Two Kinds of Explanation. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pages 819–824.

Thagard, P. and Stewart, T. C. (2011). The aha! experience: Creativity through emergent binding in neural networks. *Cognitive Science*, 35(1):1–33.

Tredennick, H., editor (1933). *Metaphysics.* Loeb Classical Library. Harvard University Press.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59:433–460.

Turing, A. (1969). Intelligent machinery. In Meltzer, B. and Michie, D., editors, *Machine Intelligence*, volume 5, pages 3–23. Edinburgh University Press.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, 90(4):293–315.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, 32:939–984.

Vazirani, V. (2010). *Approximation Algorithms.* Springer.

Veale, T. and O'Donoghue, D. (2000). Computation and blending. *Cognitive Linguistics*, 11(3/4):253–281.

Wareham, T., Kwisthout, J., Haselager, P., and van Rooij, I. (2011). Ignorance is bliss: A complexity perspective on adapting reactive architectures. In *Proceedings of the First IEEE Conference on Development and Learning and on Epigenetic Robotics*, pages 465–470.

Winograd, T. (1971). *MIT AI Technical Report 235: Procedures as a Representation for Data in a Computer Program for Understanding Natural Language.*

Winston, P. H. (1980). Learning and reasoning by analogy. *Commun. ACM*, 23(12):689–703.

Zook, K. and Maier, J. (1994). Systematic analysis of variables that contribute to the formation of analogical misconceptions. *Journal of Educational Psychology*, 86:589–600.

Part IV.

Referenced Published Articles

Chapter 3

Complex Analogies: Remarks on the Complexity of HDTP *(R. Robere and T. R. Besold)*

Abstract:

After an introduction to Heuristic-Driven Theory Projection (HDTP) as framework for computational analogy-making, and a compact primer on parametrized complexity theory, we provide a complexity analysis of the key mechanisms underlying HDTP, together with a short discussion of and reflection on the obtained results. Amongst others, we show that restricted higher-order anti-unification as currently used in HDTP is W[1]-hard (and thus NP-hard) already for quite simple cases. Also, we obtain W[2]-hardness, and NP-completeness, for the original mechanism used for reducing second-order to first-order anti-unifications in the basic version of the HDTP system.

Originally published as:

R. Robere and T. R. Besold. ***Complex Analogies: Remarks on the Complexity of HDTP***. In Proceedings of the 25th Australasian Joint Conference on Artificial Intelligence (AI 2012), LNCS Vol. 7691. Springer, 2012

URL:
`http://link.springer.com/chapter/10.1007/978-3-642-35101-3_45`

When Almost Is Not Even Close: Remarks on the Approximability of HDTP *(T. R. Besold and R. Robere)*

Abstract:

A growing number of researchers in Cognitive Science advocate the thesis that human cognitive capacities are constrained by computational tractability. If right, this thesis also can be expected to have far-reaching consequences for work in Artificial General Intelligence: Models and systems considered as basis for the development of general cognitive architectures with human-like performance would also have to comply with tractability constraints, making in-depth complexity theoretic analysis a necessary and important part of the standard research and development cycle already from a rather early stage. In this paper we present an application case study for such an analysis based on results from a parametrized complexity and approximation theoretic analysis of the Heuristic Driven Theory Projection (HDTP) analogy-making framework.

Originally published as:

T. R. Besold and R. Robere. ***When Almost Is Not Even Close: Remarks on the Approximability of HDTP***. In Artificial General Intelligence - 6th International Conference (AGI 2013), LNCS Vol. 7999. Springer, 2013

URL:
`http://link.springer.com/chapter/10.1007/978-3-642-39521-5_2`

Chapter 4

Towards a Domain-Independent Computational Framework for Theory Blending *(M. Martinez et al.)*

Abstract:

The literature on conceptual blending and metaphor-making has illustrations galore of how these mechanisms may support the creation and grounding of new concepts (or whole domains) in terms of a complex, integrated network of older ones. In spite of this, as of yet there is no general computational account of blending and metaphor-making that has proven powerful enough as to cover all the examples from the literature. This paper proposes a logic-based framework for blending and metaphor making and explores its applicability in settings as diverse as mathematical domain formation, classical rationality puzzles, and noun-noun combinations.

Originally published as:

M. Martinez, T. R. Besold, A. Abdel-Fattah, K.-U. Kühnberger, H. Gust, M. Schmidt, and U. Krumnack. ***Towards a domain-independent computational framework for theory blending***. In AAAI Technical Report of the AAAI Fall 2011 Symposium on Advances in Cognitive Systems. AAAI Press, 2011

URL:
`https://www.aaai.org/ocs/index.php/FSS/FSS11/paper/view/4144`

Theory Blending as Framework for Creativity in Systems for General Intelligence *(M. Martinez et al.)*

Abstract:

Being creative is a central property of humans in solving problems, adapting to new states of affairs, applying successful strategies in previously unseen situations, or coming up with new conceptualizations. General intelligent systems should have a potential to realize to a certain extent such forms of creativity. We think that creativity and productivity issues can be best addressed by taking cognitive mechanisms into account, such as analogy–making, concept blending, computing generalizations and the like. In this paper, we argue for the usage of such mechanisms for creativity. We exemplify the potential of such a mechanisms like theory blending using a historical example from mathematics. Furthermore, we argue for the claim that modeling creativity by such mechanisms has a huge potential in a variety of domains.

Originally published as:

M. Martinez, T. R. Besold, A. Abdel-Fattah, H. Gust, M. Schmidt, U. Krumnack and K.-U. Kühnberger. Foundations of Artificial General Intelligence, chapter ***Theory Blending as a Framework for Creativity in Systems for General Intelligence***. Thinking Machines. Studies in Computational Cognition. Atlantis - Springer, 2012

URL:
`http://link.springer.com/chapter/10.2991/978-94-91216-62-6_12`

Chapter 5

Analogy Engines in Classroom Teaching: Modeling the String Circuit Analogy *(T. R. Besold)*

Abstract:

The importance of analogy-making and analogy-based reasoning for human cognition and learning by now has widely been recognized, and analogy-based methods are slowly also being explicitly integrated into the canon of approved education and teaching techniques. Still, the actual level of knowledge about analogy as instructional means and device as of today is rather low and subject to scientific study and investigation. In this paper, we propose the fruitful use of computational analogy-engines as methodological tool in this domain of research, motivating our claim by a short case study showing how Heuristic-Driven Theory Projection can be used to model the mode of operation of an analogy taken from a science class for 8 to 9 year old children.

Originally published as:

T. R. Besold. ***Analogy Engines in Classroom Teaching: Modeling the String Circuit Analogy***. In AAAI Technical Report of the AAAI Spring 2013 Symposium on Creativity and (Early) Cognitive Development. AAAI Press, 2013

URL:
`http://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5720`

Analogy and Arithmetic: An HDTP-Based Model of the Calculation Circular Staircase *(T. R. Besold et al.)*

Abstract:

Analogical reasoning and its applications are gaining attention not only in cognitive science but also in the context of education and teaching. In this paper we provide a short analysis and a detailed formal model (based on the Heuristic-Driven Theory Projection framework for computational analogy-making) of the Calculation Circular Staircase, a tool for teaching basic arithmetic and insights based on the ordinal number conception of the natural numbers to children in their first years of primary school. We argue that such formal methods and computational accounts of analogy-making can be used to gain additional insights in the inner workings of analogy-based educational methods and tools.

Originally published as:

T. R. Besold, A. Pease and M. Schmidt. ***Analogy and Arithmetic: An HDTP-Based Model of the Calculation Circular Staircase***. In Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society, 2013

URL:
`http://mindmodeling.org/cogsci2013/papers/0351/index.html`

Applying AI for Modeling and Understanding Analogy-Based Classroom Teaching Tools & Techniques *(T. R Besold and K.-U. Kühnberger)*

Abstract:

This paper forms the final part of a short series of related articles dedicated to highlighting a fruitful type of application of cognitively-inspired analogy engines in an educational context. It complements the earlier work with an additional fully worked out example by providing a short analysis and a detailed formal model (based on the Heuristic-Driven Theory Projection computational analogy framework) of the Number Highrise, a tool for teaching multiplication-based relations in the range of natural numbers up to 100 to children in their first years of primary school.

Originally published as:

T. R. Besold and K.-U. Kühnberger. ***Applying AI for Modeling and Understanding Analogy-Based Classroom Teaching Tools & Techniques***. In C. Lutz and M. Thielscher, editors, KI 2014: Advances in Artificial Intelligence, Proceedings of the 37th Annual German Conference on AI, LNCS Vol. 8736. Springer, 2014

URL:
http://link.springer.com/chapter/10.1007/978-3-319-11206-0_6

Sensorimotor Analogies in Learning Abstract Skills and Knowledge: Modeling Analogy-Supported Education in Mathematics and Physics *(T. R. Besold)*

Abstract:

In this summary report I give an account of research conducted over the last two years, showing the suitability and the advantages of applying computational analogy-engines in the analysis and design of analogy-based methods and tools in teaching and education. This overview constitutes the conclusion of the first phase of a multi-stage effort trying to introduce computational models of analogy also to education and the learning sciences, thus opening up these fields to computational tools and methods not only on an instrumental level, but also in analytical, conceptual, and design-oriented studies. I locate the "analogy-engines in the classroom" research program within the bigger schemes of studying human creativity and computational creativity, provide an introduction to the theoretical underpinnings of the endeavor, and revisit three worked out case studies serving as proofs of the feasibility of the overall approach.

Originally published as:

T. R. Besold. ***Sensorimotor Analogies in Learning Abstract Skills and Knowledge: Modeling Analogy-Supported Education in Mathematics and Physics***. In AAAI Technical Report of the AAAI Fall 2014 Symposium on Modeling Changing Perspectives: Reconceptualizing Sensorimotor Experiences. AAAI Press, 2014

URL:
`http://www.aaai.org/Press/Reports/Symposia/Fall/fs-14-05.php`

Chapter 6

Rationality {in|for|through} AI *(T. R. Besold)*

Abstract:

Based on an assessment of the history and status quo of the concept of rationality within AI, I propose to establish research on (artificial) rationality as a research program in its own right, aiming at developing appropriate notions and theories of rationality suitable for the special needs and purposes of AI. I identify already existing initial attempts at and possible foundations of such an endeavor, give an account of motivations, expected consequences and rewards, and outline how such a program could be linked to efforts in other disciplines.

Originally published as:

T. R. Besold. ***Rationality in/for/through AI***. In Jozef Kelemen, Jan Romportl, and Eva Zackova, editors, Beyond Artificial Intelligence: Contemplations, Expectations, Applications. Topics in Intelligent Engineering and Informatics, vol. 4. Springer, 2013

URL:
`http://link.springer.com/chapter/10.1007/978-3-642-34422-0_3`

E Pluribus Multa in Unum: The Rationality Multiverse *(T. R. Besold and K.-U. Kühnberger)*

Abstract:

The paper argues for a new view on and an approach to rationality as a concept of study and modeling paradigm of human behavior. After critically reviewing classical (normative) approaches to rationality, decision-making, and rational behavior, we present cornerstones of a positive, integrative, and holistic conception of these cognitive capacities. A discussion of key elements of this new view is given, and possible consequences and implications are considered.

Originally published as:

T. R. Besold and K.-U. Kühnberger. ***E Pluribus Multa In Unum: The Rationality Multiverse***. In N. Miyake, D. Peebles and R. P. Cooper, editors, Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society, 2012

URL:
`http://mindmodeling.org/cogsci2012/papers/0238/index.html`

Rationality in Context: An Analogical Perspective *(T. R. Besold)*

Abstract:

At times, human behavior seems erratic and irrational. Therefore, when modeling human decision-making, it seems reasonable to take the remarkable abilities of humans into account with respect to rational behavior, but also their apparent deviations from the normative standards of rationality shining up in certain rationality tasks. Based on well-known challenges for human rationality, together with results from psychological studies on decision-making and from previous work in the field of computational modeling of analogy-making, I argue that the analysis and modeling of rational belief and behavior should also consider context-related cognitive mechanisms like analogy-making and coherence maximization of the background theory. Subsequently, I conceptually outline a high-level algorithmic approach for a Heuristic Driven Theory Projection-based system for simulating context-dependent human-style rational behavior. Finally, I show and elaborate on the close connections, but also on the significant differences, of this approach to notions of "ecological rationality".

Originally published as:

T. R. Besold. ***Rationality in Context: An Analogical Perspective***. In Modeling and Using Context, LNCS Vol. 8175. Springer, 2013

URL:
`http://link.springer.com/chapter/10.1007/978-3-642-40972-1_10`

Chapter 7

When Thinking Never Comes to a Halt: Using Formal Methods in Making Sure Your AI Gets the Job Done Good Enough *(T. R. Besold and R. Robere)*

Abstract:

The recognition that human minds/brains are finite systems with limited resources for computation has led researchers in cognitive science to advance the Tractable Cognition thesis: Human cognitive capacities are constrained by computational tractability. As also human-level AI in its attempt to recreate intelligence and capacities inspired by the human mind is dealing with finite systems, transferring this thesis and adapting it accordingly may give rise to insights that can help in progressing towards meeting the classical goal of AI in creating machines equipped with capacities rivaling human intelligence. Therefore, we develop the "Tractable Artificial and General Intelligence Thesis" and corresponding formal models usable for guiding the development of cognitive systems and models by applying notions from parameterized complexity theory and the theory of hardness of approximation to a general AI framework. In this chapter we provide an overview of our work, putting special emphasis on connections and correspondences to the heuristics framework as recent development within cognitive science and cognitive psychology.

Originally published as:

T. R. Besold and R. Robere (forthcoming). ***When Thinking Never Comes to a Halt: Using Formal Methods in Making Sure Your AI Gets the Job Done Good Enough***. In V. C. Müller (Ed.), Fundamental Issues of Artificial Intelligence (Synthese Library). Springer, 2015

URL:
`(forthcoming)`

When Thinking Never Comes to a Halt: Using Formal Methods in Making Sure Your AI Gets the Job Done Good Enough*

Tarek R. Besold and Robert Robere

Abstract The recognition that human minds/brains are finite systems with limited resources for computation has led researchers in cognitive science to advance the Tractable Cognition thesis: Human cognitive capacities are constrained by computational tractability. As also human-level AI in its attempt to recreate intelligence and capacities inspired by the human mind is dealing with finite systems, transferring this thesis and adapting it accordingly may give rise to insights that can help in progressing towards meeting the classical goal of AI in creating machines equipped with capacities rivaling human intelligence. Therefore, we develop the "Tractable Artificial and General Intelligence Thesis" and corresponding formal models usable for guiding the development of cognitive systems and models by applying notions from parameterized complexity theory and hardness of approximation to a general AI framework. In this chapter we provide an overview of our work, putting special emphasis on connections and correspondences to the heuristics framework as recent development within cognitive science and cognitive psychology.

1 Introduction: The Importance of Formal Analysis for Cognitive Systems Research

After a certain abandonment of the original dream(s) of artificial intelligence (AI) towards the end of the last century, research in cognitive systems, artificial human-level intelligence, complex cognition, and integrated intelligent systems over the last decade has witnessed a revival and is now entering its second spring with several specifically dedicated conference series, symposia, workshops, journals and a growing number of books and high-profile research projects. Still, quite some fundamental questions remain to be answered before a unified approach to solving

Tarek R. Besold
Institute of Cognitive Science, University of Osnabrück, Germany
e-mail: tbesold@uni-osnabrueck.de

Robert Robere
Department of Computer Science, University of Toronto, Canada
e-mail: robere@cs.toronto.edu

* Besold, Tarek R. and Robere, Robert (forthcoming in 2015), 'When Thinking Never Comes to a Halt: Using Formal Methods in Making Sure Your AI Gets the Job Done Good Enough', in V. C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Synthese Library; Berlin: Springer).

the big riddles underlying the (re)creation of human-level intelligence and cognition may arise. Currently, there are many different paradigms competing with each other: Symbolism versus connectionism, high-level modeling of specific cognitive capacities versus low-level models with emergent behavior, holistic versus modular approaches.

Each of these paradigms brings along its own terminology, conceptual perspective, and engineering methods, resulting in a wide variety of approaches to solving the intelligence puzzle. This, in turn, makes it hard to establish standards and insights in cognitive models and cognitive systems which are valid on a general level independent of the chosen perspective and methodology. Still, there are a few elements common to most (if not all) of the mentioned approaches (in that, for instance, they are applied in attempts to model one or several human cognitive capacities), making the wish for general principles and results more urgent. Here, formal methods and analysis can provide a solution: Due to their general nature they can often be applied without prior commitment to a particular formalism or architecture, allowing to establish high-level insights and generally applicable findings. In other words, these techniques can provide guidelines and hints at how to unify approaches and progress towards the overall goals of the respective research programs.

In what follows, we give an overview of the status quo of our work on the topic, combining previous independently published contributions and extending the individual pieces into a unified whole. This summary shall provide both evidence supporting the just made claims about the possible role of formal methods for general high-level AI design, and concrete insights concerning heuristics and their use in cognitive systems as important specific example. Sect. 2 introduces the mindset underlying our work before Sect. 3 summarizes important theoretical results, followed by a worked application case for our approach in Sect. 4. Opening the second half of the chapter, Sect. 5 then elaborates the connection between the notion of cognitive heuristics (and their models) to recent results from parameterized complexity and the theory of hardness of approximation, before Sect. 6 addresses some of the most common criticisms targeting the application of formal methods to work in AI and cognitive systems. Sect. 7 concludes the chapter, connecting it to related work by other scholars and pointing out some future directions of development.

2 Complexity and Cognition

Two famous ideas conceptually lie at the heart of many endeavors in computational cognitive modeling, cognitive systems research and artificial intelligence: The "computer metaphor" of the mind, i.e. the concept of a computational theory of mind as described in Pylyshyn (1980), and the Church-Turing thesis (a familiar version of which is stated in Turing (1969)). The former bridges the gap between humans and computers by advocating the claim that the human mind and brain can be seen as an information processing system and that reasoning and thinking corresponds to processes that meet the technical definition of computation as formal symbol manipulation, the latter gives an account of the nature and limitations of the computational power of such a system: Every function for which there is an algorithm (i.e., those functions which are "intuitively computable" by a sequence of steps) is computable by a Turing machine — and functions that are not computable by such a machine are to be considered not computable in principle by any machine.

But the "computer metaphor" and the Church-Turing thesis also had significant impact on cognitive science and cognitive psychology. As stated in Cummins (2000), one of the primary aims of cognitive psychology is to explain human cognitive capacities — which are often modeled in terms of computational-level theories of cognitive processes (i.e., as precise characterizations of the hypothesized inputs and outputs of the respective capacities together with the functional mappings between them; c.f. Marr (1982) for details). Unfortunately, computational-level theories are often underconstrained by the available empirical data, allowing for several different input-output mappings and corresponding theories. A first attempt at mitigating this problem can now be based on the aforegiven Church-Turing thesis: If the thesis were true, the set of functions computable by a cognitive system would be a subset of the Turing-computable functions. Now, if a computational-level theory would assume that the cognitive system under study computes a function uncomputable by a Turing machine then the model could already be rejected on theoretical grounds.

Still, directly applying the notion of Turing computability as equivalent to the power of cognitive computation has to be considered overly simplistic. Whilst it has been commonly accepted that the thesis holds in general, its practical relevance in cognitive agents and systems is at least questionable. As already recognized in Simon (1957), actual cognitive systems (due to their nature as physical systems) need to perform their tasks in limited time and with a limited amount of space at their disposal. Therefore, the Church-Turing thesis by itself is not strict enough for being used as a constraint on cognitive theories as it does not take into account any of these dimensions. To mitigate this problem, different researchers over the last decades have proposed the use of mathematical complexity theory, like the concept of NP-completeness, as an assisting tool (see, e.g., Levesque (1988); Frixione (2001)), bringing forth the so called "*P-Cognition thesis*": Human cognitive capacities are hypothesized to be of the polynomial-time computable type.

However, using the "polynomial-time computable" as synonymous with "efficient" may already be overly restrictive. In modern times there are many examples of problems which have algorithms that have worst-case exponential behaviour, but tend to work quite well in practice on small inputs (take, for example, any of the modern algorithms for the travelling salesperson problem). However, this is not the type of restriction that we will focus on. Instead we take the following viewpoint: it is often the case that we as humans are able to solve problems which may be hard in general but suddenly become feasible if certain parameters of the problem are restricted. This idea has been formalized in the field of *parameterized complexity theory*, in which "tractability" is captured by the class of fixed-parameter tractable problems FPT:[2]

Definition 1. FPT
A problem P is in FPT if P admits an $O(f(\kappa)n^c)$ algorithm, where n is the input size, κ is a parameter of the input constrained to be "small", c is an independent constant, and f is some computable function.

[2] For an introduction to parameterized complexity theory see, e.g., Flum and Grohe (2006); Downey and Fellows (1999).

Originating from this line of thought, van Rooij (2008) introduces a specific version of the claim that cognition and cognitive capacities are constrained by the fact that humans basically are finite systems with only limited resources for computation: Applying the just presented definition from parameterized complexity theory, this basic notion of resource-bounded computation for cognition is formalized in terms of the so called "FPT-*Cognition thesis*", demanding for human cognitive capacities to be fixed-parameter tractable for one or more input parameters that are small in practice (i.e., stating that the computational-level theories have to be in FPT).

3 Theoretical Foundation: The Tractable AGI Thesis

But whilst the aforementioned P-Cognition thesis also found its way into AI (cf., e.g., Cooper (1990); Nebel (1996)), the FPT-Cognition thesis this far has widely been ignored. Recognizing this as a serious deficit, for example in Robere and Besold (2012) and Besold and Robere (2013a), we proposed a way of (re)introducing the idea of tractable computability for cognition into AI and cognitive systems research by rephrasing and accordingly adapting the FPT-form of the Tractable Cognition thesis. As all of the currently available computing systems used for implementing cognitive models and cognitive systems are ultimately finite systems with limited resources (and thus in this respect are not different from other cognitive agents and human minds and/or brains), in close analogy we developed the "Tractable AGI thesis" (Tractable Artificial and General Intelligence thesis).

Tractable AGI thesis
Models of cognitive capacities in artificial intelligence and computational cognitive systems have to be fixed-parameter tractable for one or more input parameters that are small in practice (i.e., have to be in FPT).

Concerning the interpretation of this thesis, suppose a cognitive modeler or AI system designer is able to prove that his model at hand is — although in its most general form possibly NP-hard — fixed-parameter tractable for some set of parameters κ. This implies that if the parameters in κ are fixed small constants for problem instances realized in practice, then it is possible to efficiently compute a solution.

4 Worked Example: Complex Analogies in HDTP

Before further continuing the theoretical line of work in Sect. 5, we want to spend some time on a worked application case of analyzing a cognitive AI system by means of formal methods. We therefore give a fairly detailed reproduction of results from a parameterized complexity study of the Heuristic-Driven Theory Projection (HDTP) computational analogy-making framework (originally presented in Robere and Besold (2012)). By this we hope to show how the mostly academic-theoretical considerations from Sect. 2 and 3 directly connect to everyday AI and cognitive systems practice.

4.1 The Motivation Behind It

During the course of a day, we use different kinds of reasoning processes: We solve puzzles, play instruments, or discuss problems. Often we will find ourselves in places and times in which we apply our knowledge of a familiar situation to the (structurally similar) novel one. Today it is undoubted that one of the basic elements of human cognition is the ability to see two a priori distinct domains as similar based on their shared relational structure (i.e., analogy-making). Some prominent cognitive scientists as, for example, Hofstadter (2001), go as far as to consider analogy the core of cognition itself. Key abilities within everyday life, such as communication, social interaction, tool use, and the handling of previously unseen situations crucially rely on the use of analogy-based strategies and procedures. One of the key mechanisms underlying analogy-making, relational matching, is also the basis of perception, language, learning, memory and thinking, i.e., the constituent elements of most conceptions of cognition (Schwering et al (2009b)).

Because of this crucial role of analogy in human cognition, researchers in cognitive science and artificial intelligence have been creating computational models of analogy-making since the advent of computer systems. But the field has changed significantly during that time: Early work such as that of Reitman et al (1964) or Evans (1964) should serve as a proof of concept for the possibilities and the power of AI systems, possibly paving the way for more flexible approaches to reasoning and artificial cognition. Still, from a theoretical and methodological point of view these systems were not necessarily committed to considerations concerning cognitive adequacy or psychological plausibility and did not correspond to a fully developed underlying theoretical paradigm about human analogy-making. In contrast, modern analogy systems – the most prominent of which probably is the Structure-Mapping Engine (SME, Falkenhainer et al (1989)) and MAC/FAC (Gentner and Forbus (1991)) – come with their respective theory about how analogy-making works on a human scale (see, e.g., Gentner (1983) for the theory behind SME) and often even make claims not only on a computational level of description, but even hypothesize more or less precisely specified algorithmic mechanisms of analogy.

And this now is where our proposed approach for formal analysis comes into play: If human-likeness in a system's theoretical foundations and behavior is assumed, it should also automatically become clear that the same standards of evaluation and the same formal properties which are true for human cognition have to be met and have to hold for the system. Moreover this has to be the case in general and not only on a selected subset of examples or under positively limiting conditions and a priori assumptions on the possible cases a system might encounter (unless, of course, these assumptions can also be made without loss of generality for the human counterpart). Analogy-making is a prime example for this setting as — precisely due to the aforementioned variety of occurrences and manifestations of this cognitive capacity — the architect of a cognitive system model of analogy has to make sure that certain properties of the system hold true with a high degree of independence from the specific problem case at hand.

Furthermore, from a tractability perspective, computational analogy systems are a first-rate application area for our theoretical paradigm. Analogy-making has gained attention in cognitive science and cognitive AI not only because of its general applicability but also due to the fact that humans seem to be able to retrieve and use analogies in very efficient ways: Conversations happen in real time, social interaction — although highly diverse in its different levels — is pervasive and in most cases does not require attention or conscious thought (even if we are only acquainted with the general rules and paradigm and not with the specific situations we encounter), and once we understood the solution to a certain riddle or problem we have no problem immediately applying it to analogical cases even when they are completely different in appearance or setting.

4.2 The Formal Analysis

Heuristic-Driven Theory Projection, introduced in Schwering et al (2009a), is a formal theory and corresponding software implementation, conceived as a mathematically sound framework for analogy-making. HDTP has been created for computing analogical relations and inferences for domains which are given in the form of a many-sorted first-order logic representation. Source and target of the analogy-making process are defined in terms of axiomatizations, i.e., given by a finite set of formulae. HDTP tries to produce a generalization of both domains by aligning pairs of formulae from the two domains by means of a process called *anti-unification*, which tries to solve the problem of generalizing terms in a meaningful way, yielding for each term an "anti-instance" in which some subterms have been replaced by variables (which in turn would allow for a retrieval of the original terms by a substitution of the variables by appropriate subterms).

HDTP in its present version uses a restricted form of higher-order anti-unification presented in Krumnack et al (2007). In higher-order anti-unification, classical first-order terms are extended by the introduction of variables which may take arguments (where classical first-order variables correspond to variables with arity 0), making a term either a first-order or a higher-order term. Then, anti-unification can be applied analogously to the original first-order case, yielding a generalization subsuming the specific terms. The class of substitutions which are applicable in HDTP is restricted to (compositions of) the following four cases: renamings (replacing a variable by another variable of the same argument structure), fixations (replacing a variable by a function symbol of the same argument structure), argument insertions, and permutations (an operation rearranging the arguments of a term).

This formalism has proven capable of detecting structural commonalities not accessible to first-order anti-unification, as for instance also structural commonalities between functions and predicates within the logical language can be found and exploited (whilst the first-order formalism would in these be limited to the respective argument positions only), allowing for a more general recognition of relational mappings (as opposed to mere attribute mappings). Once the generalization has been computed, the alignments of formulae together with the respective generalizations can be read as proposals of analogical relations between source and target domain, and can be used for guiding an analogy-based process of transferring knowledge between both domains (see Fig. 1 for an overview of the analogy-making process). Analogical transfer results in structure enrichment on the target side, which corresponds to the addition of new axioms to the target theory, but may also involve the addition of new first-order symbols.

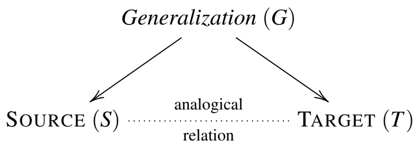

Fig. 1 A schematic overview of HDTP's generalization-based approach to analogy.

Whilst HDTP undoubtedly exhibits pleasant properties – say, in terms of expressivity of the modeling language, and clarity of the underlying conceptual approach – until recently there had been no detailed analysis of its computational tractability. In order to change this unsatisfactory state of affairs, we decided to apply some techniques from parameterized complexity theory to the system trying to better understand its strengths and weaknesses.

As already mentioned, a restricted higher-order anti-unification is defined as any composition of a certain set of unit substitutions which can formally be specified in the following way:

Definition 2. Restricted higher-order anti-unification
The following are the types of unit substitutions allowed in restricted higher-order anti-unification.

1. A renaming $\rho(F,F')$ replaces a variable $F \in \mathcal{V}_n$ with another variable $F' \in \mathcal{V}_n$:
   $F(t_1,\ldots,t_n) \xrightarrow{\rho(F,F')} F'(t_1,\ldots,t_n)$.
2. A fixation $\phi(F,f)$ replaces a variable $F \in \mathcal{V}_n$ with a function symbol $f \in \mathcal{C}_n$:
   $F(t_1,\ldots,t_n) \xrightarrow{\phi(F,f)} f(t_1,\ldots,t_n)$.
3. An argument insertion $\iota(F,F',V,i)$ is defined as follows, where $F \in \mathcal{V}_n, F' \in \mathcal{V}_{n-k+1}, V \in \mathcal{V}_k, i \in [n]$:
   $F(t_1,\ldots,t_n) \xrightarrow{\iota(F,F',V,i)} F'(t_1,\ldots,t_{i-1},V(t_i,\ldots,t_{i+k}),t_{i+k+1},\ldots,t_n)$.
   It "wraps" k of the subterms in a term using a k-ary variable, or can be used to insert a 0-ary variable.
4. A permutation $\pi(F,\tau)$ rearranges the arguments of a term, with $F \in \mathcal{V}_n$, $\tau : [n] \to [n]$ a bijection:
   $F(t_1,\ldots,t_n) \xrightarrow{\pi(F,\tau)} F(t_{\pi(1)},\ldots,t_{\pi(n)})$.

A *restricted substitution* is a substitution which results from the composition of any sequence of unit substitutions.

By considering different combinations of restricted substitutions we can define several different forms of higher-order anti-unification. Unfortunately, as already recognized by Krumnack et al (2007), the least general generalizer is not necessarily unique. Therefore, in our analysis we instead consider decision versions of the problems parameterized by the number of substitutions, variables, and types of variables used.

Problem 1. F Anti-Unification
Input: Two terms f,g, and a natural $k \in \mathbb{N}$
Problem: Is there an anti-unifier h, containing at least k variables, using only renamings and fixations?

Problem 2. FP Anti-Unification
Input: Two terms f, g, and naturals $l, m, p \in \mathbb{N}$.
Problem: Is there an anti-unifier h, containing at least l 0-ary variables and at least m higher arity variables, and two substitutions σ, τ using only renamings, fixations, and at most p permutations such that $h \xrightarrow{\sigma} f$ and $h \xrightarrow{\tau} g$?

Problem 3. FPA Anti-Unification
Input: Two terms f, g and naturals $l, m, p, a \in \mathbb{N}$.
Problem: Is there an anti-unifier h, containing at least l 0-ary variables, at least m higher arity variables, and two substitutions σ, τ using renamings, fixations, at most p permutations, and at most a argument insertions such that $h \xrightarrow{\sigma} f$ and $h \xrightarrow{\tau} g$?

We summarize our (parameterized) complexity-theoretic results of higher-order anti-unification in the following theorem:[3]

Theorem 1.
1.) F Anti-Unification is solvable in polynomial time.
2.) FP Anti-Unification is NP-complete and $\mathsf{W}[1]$-hard w.r.t. parameter set $\{m, p\}$.
3.) Let r be the maximum arity and s be the maximum number of subterms of the input terms. Then FP Anti-Unification is in FPT w.r.t. parameter set $\{s, r, p\}$.
4.) FPA Anti-Unification is NP-complete and $\mathsf{W}[1]$-hard w.r.t. parameter set $\{m, p, a\}$.

4.3 Interpretation of the Results

We want to provide some thoughts on the consequences of the complexity results from the previous section, putting the obtained insights into a cognitive AI context and thereby making the intrinsic connection between the formal considerations and the analogy mechanism of the implemented system explicit.

We focus on a result directly affecting HDTP. The result showing that FP higher-order anti-unification is $\mathsf{W}[1]$-hard gives a hint at the difficulty introduced by the operations admissible within the restricted higher-order anti-unification on the complexity of the analogy-making process. Indeed, the only way that FP anti-unification can restructure the order of the terms is by argument permutations, and our results show that even allowing a *single* permutation is enough to imply computational hardness. If we contrast this result against the polynomial-time algorithm for F anti-unification, we have evidence that even a slight ability to restructure the input terms makes higher-order anti-unification a difficult problem to solve.

Now, additionally making the (most likely reasonable) assumption that $\mathsf{P} \neq \mathsf{NP}$ (and $\mathsf{FPT} \neq \mathsf{W}[1]$) holds, the presented hardness results cast a shadow on the suitability of the HDTP framework in its present state as basis for a general model for high-level cognitive capacities or a general cognitive architecture. Still, it should be noticed that the mere fact that HDTP in its present state is basically intractable does not mean that future versions cannot be made tractable. Here, the insights obtained from the formal analysis can serve as guidelines for the future evolution of

[3] The corresponding proofs of the respective results can be found in Robere and Besold (2012). Moreover, in the theorem statements $\mathsf{W}[1]$ refers to the class of problems solvable by constant depth combinatorial circuits with at most 1 gate with unbounded fan-in on any path from an input gate to an output gate. In parameterized complexity, the assumption $\mathsf{W}[1] \neq \mathsf{FPT}$ can be seen as analogous to $\mathsf{P} \neq \mathsf{NP}$.

the system: In our opinion, one of the main questions for future theoretical research in relation to HDTP will have to address the question of how HDTP's version of computing generalizations via restricted higher-order anti-unification can be further constrained in a meaningful way as to obtain maximal expressivity and applicability whilst still staying within the domain of polynomial solvability. Also, more parametrized analysis will be needed, showing which are the factors that really impact complexity, and which are aspects of a problem that are not really harmful.

5 Setting Limits to Heuristics in Cognitive Systems

Leaving the basic considerations and the application study of formal means of analysis of cognitive AI systems behind us we want to return to a more theoretical part of our work. A still growing number of researchers in cognitive science and cognitive psychology, starting in the 1970s with the "heuristics and biases" program Kahneman et al (1982), and today prominently heralded, for instance, in the work of Gigerenzer and colleagues Gigerenzer et al (2011), argues that humans in their common sense reasoning do not apply any full-fledged form of logical or probabilistic reasoning to possibly highly complex problems, but instead rely on mechanisms — which are mostly automatic and unconcious — that allow them to circumvent the impending complexity explosion and nonetheless reach acceptable solutions to the original problems. Amongst the plethora of proclaimed automatisms are, for example, the representativeness heuristic Kahneman et al (1982) or the take-the-best heuristic Czerlinski et al (1999).

All of these mechanisms are commonly subsumed under the all-encompassing general term "heuristics". Still, on theoretical grounds, at least two quite different general types of approach have to be distinguished within this category: Either the complexity of solving a problem can be reduced by reducing the problem instance under consideration to a simpler (but solution equivalent) one, or the problem instance stays untouched but — instead of being perfectly (i.e., precisely) solved — is dealt with in a good enough (i.e., approximate) way.

Now, taking the perspective of an architect of a cognitive system considering to include human-inspired heuristics in his reasoning model for solving certain tasks, a crucial question quite straightforwardly arises: Which problems can actually be solved by applying heuristics — and how can the notion of heuristics be theoretically modeled on a sufficiently high level as to allow for a general description? Having a look at recent work in parameterized complexity theory and in hardness of approximation, we find that the two distinct types of heuristics naturally correspond to two well-known concepts from the respective fields. As firstly shown in Besold (2013) and reproduced in the following, this opens the way for establishing a solid theoretical basis for models of heuristics in cognitive systems.

5.1 The Reduction Perspective

In Sect. 3, the Tractable AGI thesis demanded for models of cognitive capacities in AI to be in FPT. However, there is also a non-trivial corollary that can be derived from this property: any instance of a problem in FPT can be reduced to a *problem kernel*.

Definition 3. Kernelization
Let P be a parameterized problem. A kernelization of P is an algorithm which takes an instance x of P with parameter κ and maps it in polynomial time to an instance y such that $x \in P$ if and only if $y \in P$, and the size of y is bounded by $f(\kappa)$ (f a computable function).

Theorem 2. Kernelizability Downey et al (1997)
A problem P is in FPT if and only if it is kernelizable.

This theorem on the one hand entails that any positive FPT result obtainable for the model in question essentially implies that there is a “downward reduction” for the underlying problem to some sort of smaller or less-complex instance of the same problem, which can then be solved — whilst on the other hand (assuming $\mathsf{W}[1] \neq \mathsf{FPT}$) any negative result implies that there is no such downward reduction. This equivalence forms a first connecting point to some of the different heuristics frameworks in cognitive science and cognitive psychology — and can also have important ramifications for the modeling of many cognitive capacities in computational cognitive systems.

On the one hand, from a constructive perspective, by actively considering the kernelizability of problems (or rather problem classes), we can provide inspiration and first hints at hypothesizing a specialized cognitive structure capable of computing the reduced instance of a problem, which then might allow for an efficient solving procedure. On the other hand, taking a more theoretical stance, by categorizing problems according to kernelizability (or their lack thereof) we also can establish a distinction between problem classes which are solvable by reduction-based heuristics and those which are not — and can thus already a priori decide whether a system implementing a reduction-based heuristics might generally be unable to solve a certain problem class.

From theoretical perspective the strict equivalence between FPT-membership and kernelizability of a problem is somewhat surprising. However, on practical and applied grounds, the correspondence should seem natural and, moreover, should fairly directly explicate the connection to the notion of reduction-based heuristics: If cognitive heuristics are as fast and frugal as commonly claimed, considering them anything but (at worst) polynomial-time bounded processes seems questionable. But now, if the reduced problem shall be solvable under resource-critical conditions, using the line of argument from Sect. 2 and 3, we can just hope for it to be in FPT. Now, combining the FPT-membership of the reduced problem with the polynomial-time complexity of the heuristics, already the original problem had to be fixed-parameter tractable. Still, reduction-based heuristics are not trivialized by this: Although original and reduced problem are in FPT, the respective size of the parameters may still differ between instances (which possibly can make an important difference in application scenarios for implemented cognitive systems).

5.2 The Approximation Perspective

There is also a complementary perspective offering an alternate possibility of (re)interpreting heuristics, namely the theory of approximation algorithms: Instead of precisely solving a kernel as proposed by reduction-based heuristics, compute an approximate solution to the original problem (i.e., the solution to a relaxed problem). The idea is not any more to perfectly solve the problem (or an equivalent instance of the same class), but to instead solve the problem to some "satisfactory degree".

Here, a candidate lending itself for being considered a standard analogous to FPT in the Tractable AGI thesis is APX, the class of problems allowing polynomial-time approximation algorithms:

Definition 4. APX
An optimization problem P is in APX if P admits a constant-factor approximation algorithm, i.e., there is a constant factor $\varepsilon > 0$ and an algorithm which takes an instance of P of size n and, in time polynomial in n, produces a solution that is within a factor $1+\varepsilon$ of being optimal (or $1-\varepsilon$ for maximization problems).

Clearly, here the meaningfulness and usefulness of the theoretical notion in practice crucially depends on the choice of the bounding constant for the approximation ratio: If the former is meaningfully chosen with respect to the problem at hand, constant-factor approximation allows for quantifying the "good enough" aspect of the problem solution and, thus, offers a straightforward way of modeling the notion of "satisficing" Simon (1956) (which in turn is central to many heuristics considered in cognitive science and psychology).

One should also be careful to note that ”constant-factor approximation” is also quite unrestrictive in ways other than pure tractability. While this class captures problems that may have efficient algorithms which produce an solution that is, say, half as good as an optimal solution, it also contains problems that have efficient 1/1000-approximations, or 1/1000000. While these approximation factors almost never appear in practice, they are theoretically allowed. However, we believe in principal that this serves to strengthen any *negative* results which would place problems outside of APX.

As in the case of the reduction-based heuristics, one of the main advantages of formally identifying approximation-based heuristics with APX lies in its limiting power: If a problem shall be solved at least within a certain range from the optimal solution, but it turns out that the problem does not admit constant-factor approximation for the corresponding approximation parameter, the problem can a priori be discarded as unsolvable with approximation-based heuristics (unless one wants to also admit exponential-time mechanisms, which might be useful in selected cases but for simple complexity considerations does not seem to be feasible as a general approach).[4]

[4] On the other hand, considering more restrictive notions than APX as, for instance, PTAS (the class of problems for which there exists a polynomial-time approximation scheme, i.e., an algorithm which takes an instance of a optimization problem and a parameter $\varepsilon > 0$ and, in polynomial time, solves the problem within a factor $1+\varepsilon$ of the optimal solution) does not seem meaningful to us either, as also human satisficing does not approximate optimal solutions up to an arbitrary degree but in experiments normally yields rather clearly defined cut-off points at a certain approximation level.

5.3 Joining Perspectives

Having introduced the distinction between reduction-based and approximation-based heuristics, together with proposals for a formal model of the mechanisms behind the respective class, we now want to return to a more high-level view and look at heuristics in their entirety. This is also meaningful from the perspective of the initially mentioned system architect: Instead of deciding whether he wants to solve a certain type of task applying one of the two types of heuristics and then conducting the corresponding analysis, he might just want to directly check whether the problem at hand might be solvable by any of the two paradigms. Luckily, a fairly recently introduced theoretical concept allows for the integration of the two different views — FPT and APX can both be combined via the concept of fixed-parameter approximabiltiy and the corresponding problem class FPA:

Definition 5. FPA
The fixed-parameter version P of a minimization problem is in FPA if — for a recursive function f, a constant k, and some fixed recursive function g — there exists an algorithm such that for any given problem instance I with parameter k, and question $OPT(I) \leq k$, the algorithm which runs in $O(f(k)n^c)$ (where $n = |I|$) either outputs "no" or produces a solution of cost at most $g(k)$.

As shown in Cai and Huang (2006), both polynomial-time approximability and fixed-parameter tractability with witness (see Cai and Chen (1997) for details) independently imply the more general fixed-parameter approximability. And also on interpretation level FPA artlessly combines both views of heuristics, at a time in its approximability character accommodating for the notion of satisficing and in its fixed-parameter character accounting for the possibility of complexity reduction by kernelizing whilst keeping key parameters of the problem fixed.

Clearly, the notion of fixed-parameter approximability is significantly weaker than either FPT and kernelization, or APX. Nonetheless, its two main advantages are the all-encompassing generality (independent of the type of heuristics) and yet again the introduction of a categorization over problem types: If problems of a certain kind are not in FPA, this also excludes membership in any of the two stricter classes — and thus (in accordance with the lines of argument given above) in consequence hinders solvability by either type of heuristics.

Recalling the Tractable AGI thesis introduced in Sect. 3, we can use the just outlined conception of FPA for not only considering classical strict processing and reasoning in cognitive systems, but for also accounting for models of cognitive heuristics. This allows us to adapt the original thesis into a (significantly weaker but possibly more "cognitively adequate") second form:

Fixed-Parameter Approximable AGI thesis
Models of cognitive capacities in artificial intelligence and computational cognitive systems have to be fixed-parameter approximable for one or more input parameters that are small in practice (i.e., have to be in FPA).

Whilst the original Tractable AGI thesis was aimed at AI in general (thus also including forms of high-level AI which may not be human-inspired, or which in their results shall not appear human-like), the just postulated thesis, due to its strong rooting in the (re)implementation of human-style processing, explicitly targets researchers in cognitive systems and cognitive AI.

6 The Importance of Formal Analysis for Cognitive Systems Research Revisited

An often heard fundamental criticism of trying to apply methods from complexity theory and formal computational analysis to cognitive systems and cognitive models are variations of the claim that there is no reason to characterize human behavior in terms of some "problem" (i.e., in terms of a well-defined class of computational tasks), from this drawing the conclusion that computational complexity is just irrelevant for the respective topics and fields.[5] In most cases, this judgement seems to be based on either a misconception of what a computational-level theory in the sense of Marr (1982) is (which we will refer to as "*description error*"), or a misunderstanding of what kind of claim complexity theory makes on this type of theory (in the following referred to as "*interpretation error*").

The *description error* basically questions the possibility of describing human behavior in terms of classes of computational tasks. The corresponding argument is mostly based on the perceived enormous differences between distinct manifestations of one and the same cognitive capacity already within a single subject (at the moment for descriptive simplicity's sake — without loss of generality — leaving aside the seemingly even more hopeless case of several subjects). Still, precisely here Marr's Tri-Level Hypothesis (i.e., the idea that information processing systems should be analyzed and understood at three different — though interlinked — levels) comes into play. Marr proposed three levels of description for a (biological) cognitive system, namely a computational, an algorithmic, and an implementation level. Whilst the latter two are concerned with how a system does what it does from a procedural and representational perspective (algorithmic level), and how a system is physically realized (implementation level), the computational level takes the most abstract point of view in asking for a description of what the system does in terms of giving a function mapping certain inputs on corresponding outputs. So what is needed to specify a computational-level theory of a cognitive capacity[6] is just a specified set of inputs, a set of corresponding outputs (each output corresponding to at least one element from the set of inputs) and a function establishing the connection between both (i.e., mapping each input onto an output). But now, due to the high degree of abstraction of the descriptive level, this allows us to characterize human cognitive capacities in general in terms of a computational-level theory by specifying the aforementioned three elements — where inputs and outputs are normally provided (and thus defined) by generalization from the real world environment, and the function has to be hypothesized by the respective researcher. [7] Once

[5] For reasons unclear to the authors this perspective seems to be more widespread and far deeper rooted in AI and cognitive systems research than in (theoretical) cognitive science and cognitive modeling where complexity analysis and formal computational analysis in general by now have gained a solid foothold.

[6] Here we presuppose that cognitive capacities can be seen as information processing systems. Still, this seems to be a fairly unproblematic claim, as it simply aligns cognitive processes with computations processing incoming information (e.g., from sensory input) and resulting in a certain output (e.g., a certain behavioral or mental reaction) dependent on the input.

[7] Fortunately, this way of conceptualizing a cognitive capacity naturally links to research in artificial cognitive systems. When trying to build a system modeling one or several selected cognitive capacities, we consider a general set of inputs (namely all scenarios in which a manifestation of the

all three parts have been defined, formal computational analyses can directly be conducted on the obtained computational-level theory as the latter happens to coincide in form with the type of problem (i.e., formal definition of a class of computational tasks) studied in computational complexity and approximation theory. And also the existence of at least one computational-level theory for each cognitive capacity is guaranteed: Simply take the sets of possible inputs and corresponding outputs, and define the mapping function element-wise on pairs of elements from the input and output, basically creating a lookup table returning for each possible input the respective output.

Leaving the description error behind us, we want to have a look at the *interpretation error* as further common misconception. Even when modifying the initial criticism by not questioning the overall possibility of characterizing human behavior in terms of classes of computational tasks, but rather by stating that even if there were these classes, it would not have to be the case that humans have to be able to solve all instances of a problem within a particular class, we believe that this argument is missing the point. First and foremost, as further elaborated upon in the initial paragraph of the following section, we propose to use complexity and (in)-approximability results rather as a safeguard and guideline than as an absolute exclusion criterion: As long as a computational-level theory underlying a computational cognitive model is in FPT, APX, or FPA — where in each case the system architect has to decide which standard(s) to use — the modeler can be sure that his model will do well in terms of performance for whatever instance of the problem it will encounter.[8] Furthermore, it is clear that in cognitive systems and cognitive models in general a worst-case complexity or approximability analysis for a certain problem class only rarely (if at all) can be taken as an absolute disqualifier for the corresponding computational-level theory.[9] It might well be the case that the majority of problem instances within the respective class is found to be well behaving and easily solvable, whilst the number of worst-case instances is very limited (and thus possibly unlikely to be encountered on a basis frequent enough as to turn their occurrence into a problem). However, at a high level, complexity theory can still provided researchers with meaningful information — given a computational

cognitive capacity can occur) which we necessarily formally characterize — although maybe only implicitly — in order to make the input parsable for the system, hypothesize a function mapping inputs onto outputs (namely the computations we have the system apply to the inputs) and finally obtain a well-characterized set of outputs (namely all the outputs our system can produce given its programming and the set of inputs).

[8] In discussions with researchers working in AI and cognitive systems very occasionally critical feedback relating to the choice of FPT, APX, and FPA as reference classes has been given, as these have (curiously enough) been perceived as too less restrictive. Harshly contrasting with the previously discussed criticism it was argued that human-level cognitive processing should be of linear complexity or less. Still, we do not see a problem here: Neither are we fundamentalist about this precise choice of upper boundaries, nor do we claim that these are the only meaningfully applicable ones. Nonetheless, we decided for them because they can quite straightforwardly be justified and are backed up by close correspondences with other relevant notions from theoretical and practical studies in cognitive science and AI.

[9] Of course this also explicitly includes the case in which the considered classes are conceptually not restricted to the rather coarse-grained hierarchy used in "traditional" complexity theory, but if also the significantly finer and more subtle possibilities of class definition and differentiation introduced by parametrized complexity theory and other recent developments are taken into account.

intractability or inapproximability result the researcher has an opportunity to refocus their energies onto algorithms or analysis which are more likely to be fruitful. Moreover, in the process of the formal analysis, the researcher now becomes more intimately familiar with the problem at hand — which parameters of the problem are responsible for a ”complexity explosion”, which paramaters can be allowed to grow in an unbounded fashion and still maintain computational efficiency. And, of course, if one is still put off by this sort of complexity analysis, it may simply be a matter of changing the particular analysis type: Where worst-case analyses may on certain grounds be questionable as decisive criterion about the overall usefulness of a particular computational-level theory for a cognitive capacity, average-case analyses (which admittedly are significantly harder to perform) can change the picture dramatically.

7 Conclusion: Limiting the Limits

A second frequent criticism (besides the popular general objection discussed in the previous section) against the type of work presented in this paper is that demanding for cognitive systems and models to work within certain complexity limits might always be overly restrictive: Maybe each and every human mental activity actually is performed as an exponential-time procedure, but this is never noticed as the exponent for some reason always stays very small. Undoubtedly, using what we just presented, we cannot exclude this possibility — but this also is not our current aim. What we want to say is different: We do not claim that cognitive processes are without exception within FPT, APX, or FPA, but we maintain that as long as cognitive systems and models stay within these boundaries they can safely be assumed to be plausible candidates for application in a resource-bounded general-purpose cognitive agent (guaranteeing a high degree of generalizability, scalability, and reliability). Thus, if a system architect has good reasons for plausibly assuming that a particular type of problem in all relevant cases only appears with a small exponent for a certain exponential-time solving algorithm, it may be reasonable to just use this particular algorithm in the system. But if the architect should wonder whether a problem class is likely to be solvable by a resource-bounded human-style system in general, or if it should better be addressed using reduction-based or approximation-based heuristics, then we highly recommend to consider the lines of argument presented in the previous sections.

Concerning related work, besides the conceptually and methodologically closely related, but in its focus different efforts by van Rooij van Rooij (2008) and colleagues in theoretical cognitive science (see, e.g., Kwisthout and van Rooij (2012); Blokpoel et al (2011)), of course there also is work relating fixed-parameter complexity to AI. Still, except for very few examples as, e.g., Wareham et al (2011), the applications mostly are limited to more technical or theoretical subfields of artificial intelligence (see, e.g., Gottlob and Szeider (2008) for a partial survey) and — to the best of our knowledge — this far have not been converted into a more general guiding programmatic framework for research into human-level AI and cognitive systems. To a certain extent, a laudable exception to this observation may be found in Chapman (1987), where the author presents a high-level algorithm for general purpose planning and — using formal methods similar to the ones considered above — derives general constraints for domain-independent planning under certain assumptions on the expressivity of the action representations, together with ways of avoiding the found limitations.

We therefore in our future work hope to develop the overall framework further, also showing the usefulness and applicability of the proposed methods in different worked examples from several relevant fields: The range of eligible application scenarios spans from models of epistemic reasoning and interaction, over cognitive systems in general problem-solving scenarios, down to models for particular cognitive capacities as, for example, analogy-making (see, e.g., Sect. 4 and additionally Besold and Robere (2013b) for a proof of concept).

References

Besold TR (2013) Formal Limits to Heuristics in Cognitive Systems. In: Proceedings of the Second Annual Conference on Advances in Cognitive Systems (ACS) 2013

Besold TR, Robere R (2013a) A Note on Tractability and Artificial Intelligence. In: Kühnberger KU, Rudolph S, Wang P (eds) Artificial General Intelligence - 6th International Conference, AGI 2013, Proceedings, Springer, Lecture Notes in Computer Science, vol 7999, pp 170–173

Besold TR, Robere R (2013b) When Almost Is Not Even Close: Remarks on the Approximability of HDTP. In: Kühnberger KU, Rudolph S, Wang P (eds) Artificial General Intelligence - 6th International Conference, AGI 2013, Proceedings, Springer, Lecture Notes in Computer Science, vol 7999, pp 11–20

Blokpoel M, Kwisthout J, Wareham T, Haselager P, Toni I, van Rooij, I (2011) The computational costs of recipient design and intention recognition in communication. In: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, pp 465–470

Cai L, Chen J (1997) On fixed-parameter tractability and approximability of {NP} optimization problems. Journal of Computer and System Sciences 54(3):465 – 474, DOI http://dx.doi.org/10.1006/jcss.1997.1490

Cai L, Huang X (2006) Fixed-parameter approximation: Conceptual framework and approximability results. In: Bodlaender H, Langston M (eds) Parameterized and Exact Computation, Lecture Notes in Computer Science, vol 4169, Springer Berlin Heidelberg, pp 96–108, DOI 10.1007/11847250_9

Chapman D (1987) Planning for conjunctive goals. Artificial Intelligence 32(3):333 – 377

Cooper G (1990) The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence 42:393–405

Cummins R (2000) ”How does it work?” vs. ”What are the laws?” Two conceptions of psychological explanation. In: Keil F, Wilson R (eds) Explanation and cognition, MIT Press, pp 117–145

Czerlinski J, Goldstein D, Gigerenzer G (1999) How good are simple heuristics? In: Gigerenzer G, Todd P, the ABC Group (eds) Simple Heuristics That Make Us Smart, Oxford University Press

Downey RG, Fellows MR (1999) Parameterized Complexity. Springer

Downey RG, Fellows MR, Stege U (1997) Parameterized complexity: A framework for systematically confronting computational intractability. In: Contemporary Trends in Discrete Mathematics: From DIMACS and DIMATIA to the Future, AMS

Evans TG (1964) A heuristic program to solve geometric-analogy problems. In: Proc. of the April 21-23, 1964, spring joint computer conference, ACM, New York, NY, USA, AFIPS ’64 (Spring), pp 327–338, DOI http://doi.acm.org/10.1145/1464122.1464156

Falkenhainer B, Forbus K, Gentner D (1989) The structure-mapping engine: Algorithm and examples. Artificial Intelligence 41(1):1 – 63, DOI 10.1016/0004-3702(89)90077-5

Flum J, Grohe M (2006) Parameterized Complexity Theory. Springer

Frixione M (2001) Tractable competence. Minds and Machines 11:379–397

Gentner D (1983) Structure-mapping: A theoretical framework for analogy. Cognitive Science 7(2):155–170

Gentner D, Forbus K (1991) MAC/FAC: A Model of Similarity-based Retrieval. Cognitive Science 19:141–205

Gigerenzer G, Hertwig R, Pachur T (eds) (2011) Heuristics: The Foundation of Adaptive Behavior. Oxford University Press

Gottlob G, Szeider S (2008) Fixed-parameter algorithms for artificial intelligence, constraint satisfaction and database problems. Comput J 51(3):303–325, DOI 10.1093/comjnl/bxm056

Hofstadter D (2001) Epilogue: Analogy as the core of cognition. In: Gentner D, Holyoak K, Kokinov B (eds) The Analogical Mind: Perspectives from Cognitive Science, MIT Press, Cambridge, MA, pp 499–538

Kahneman D, Slovic P, Tversky A (1982) Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press

Krumnack U, Schwering A, Gust H, Kühnberger KU (2007) Restricted higher-order anti-unification for analogy making. In: Twentieth Australian Joint Conference on Artificial Intelligence, Springer

Kwisthout J, van Rooij I (2012) Bridging the gap between theory and practice of approximate bayesian inference. In: Proceedings of the 11th International Conference on Cognitive Modeling, pp 199–204

Levesque H (1988) Logic and the complexity of reasoning. Journal of Philosophical Logic 17:355–389

Marr D (1982) Vision: A computational investigation into the human representation and processing visual information. Freeman

Nebel B (1996) Artificial intelligence: A computational perspective. In: Brewka G (ed) Principles of knowledge representation, CSLI Publications, pp 237–266

Pylyshyn Z (1980) Computation and cognition: Issues in the foundation of cognitive science. The Behavioral and Brain Sciences 3:111–132

Reitman WR, Grove RB, Shoup RG (1964) Argus: An information-processing model of thinking. Behavioral Science 9(3):270–281, DOI 10.1002/bs.3830090312

Robere R, Besold TR (2012) Complex Analogies: Remarks on the Complexity of HDTP. In: Twentyfifth Australasian Joint Conference on Artificial Intelligence, Springer, Lecture Notes in Computer Science, vol 7691, pp 530–542

van Rooij I (2008) The tractable cognition thesis. Cognitive Science 32:939–984

Schwering A, Krumnack U, Kühnberger KU, Gust H (2009a) Syntactic principles of Heuristic-Driven Theory Projection. Journal of Cognitive Systems Research 10(3):251–269

Schwering A, Kühnberger KU, Kokinov B (2009b) Analogies: Integrating multiple cognitive abilities - guest editorial. Journal of Cognitive Systems Research 10(3)

Simon HA (1956) Rational choice and the structure of the environment. Psychological Review 63:129–138, DOI 10.1037/h0042769

Simon HA (1957) Models of man: Social and rational. John Wiley & Sons, Ltd., New York

Turing A (1969) Intelligent machinery. In: Meltzer B, Michie D (eds) Machine Intelligence, Edinburgh University Press, vol 5, pp 3–23

Wareham T, Kwisthout J, Haselager P, van Rooij I (2011) Ignorance is bliss: A complexity perspective on adapting reactive architectures. In: Proceedings of the First IEEE Conference on Development and Learning and on Epigenetic Robotics, pp 465–470

Towards Formally Well-Founded Heuristics in Cognitive AI Systems *(T. R. Besold)*

Abstract:

We report on work towards the development of a framework for the application of formal methods of analysis to cognitive systems and computational models (putting special emphasis on aspects concerning the notion of heuristics in cognitive AI) and explain why this requires the development of novel theoretical methods and tools..

Originally published as:

T. R. Besold (2014). ***Towards Formally Well-Founded Heuristics in Cognitive Systems***. Cognitive Processig, 15(1 Supplement). Springer, 2014

URL:
`http://link.springer.com/article/10.1007/s10339-014-0632-2`

Chapter 8

Human-Level Artificial Intelligence Must be a Science *(T. R. Besold)*

Abstract:

Human-level artificial intelligence (HAI) surely is a special research endeavor in more than one way: The very nature of intelligence is in the first place not entirely clear, there are no criteria commonly agreed upon necessary or sufficient for the ascription of intelligence other than similarity to human performance, there is a lack of clarity concerning how to properly investigate artificial intelligence and how to proceed after the very first steps of implementing an artificially intelligent system, etc. These and similar observations have led some researchers to claim that HAI might not be a science in the normal sense and would require a different approach. Taking a recently published paper by Cassimatis as starting point, I oppose this view, giving arguments why HAI should (and even has to) conform to normal scientific standards and methods, using the approach of psychometric artificial intelligence as one of the main foundations of my position.

Originally published as:

T. R. Besold. ***Human-Level Artificial Intelligence Must Be a Science***. In Artificial General Intelligence - 6th International Conference (AGI 2013), LNCS Vol. 7999. Springer, 2013

URL:
`http://link.springer.com/chapter/10.1007/978-3-642-39521-5_19`

A Note on Chances and Limitations of Psychometric AI *(T. R. Besold)*

Abstract:

Human-level artificial intelligence (HAI) surely is a special research endeavor in more than one way: In the first place, the very nature of intelligence is not entirely clear; there are no criteria commonly agreed upon necessary or sufficient for the ascription of intelligence other than similarity to human performance (and even this criterion is open for a plethora of possible interpretations); there is a lack of clarity concerning how to properly investigate HAI and how to proceed after the very first steps of implementing an HAI system; etc. In this note I assess the ways in which the approach of Psychometric Artificial Intelligence can (and cannot) be taken as a foundation for a scientific approach to HAI.

Originally published as:

T. R. Besold. ***A Note on Chances and Limitations of Psychometric AI***. In C. Lutz and M. Thielscher, editors, KI 2014: Advances in Artificial Intelligence, Proceedings of the 37th Annual German Conference on AI, LNCS Vol. 8736. Springer, 2014

URL:
`http://link.springer.com/chapter/10.1007/978-3-319-11206-0_5`

Turing Revisited: A Cognitively-Inspired Decomposition *(T. R. Besold)*

Abstract:

After a short assessment of the idea behind the Turing Test, its actual status and the overall role it played within AI, I propose a computational cognitive modeling-inspired decomposition of the Turing test as classical "strong AI benchmark" into at least four intermediary testing scenarios: a test for natural language understanding, an evaluation of the performance in emulating human-style rationality, an assessment of creativity-related capacities, and a measure of performance on natural language production of an AI system. I also shortly reflect on advantages and disadvantages of the approach, and conclude with some hints and proposals for further work on the topic.

Originally published as:

T. R. Besold. ***Turing Revisited: A Cognitively-Inspired Decomposition***. In V. C. Müller, editor, Theory and Philosophy of Artificial Intelligence, SAPERE. Springer, 2013

URL:
`http://link.springer.com/chapter/10.1007/978-3-642-31674-6_9`