# Rationality-Guided AGI as Cognitive Systems

**Ahmed Abdel-Fattah**, **Tarek R. Besold**, **Helmar Gust**,
**Ulf Krumnack**, **Martin Schmidt**, **Kai-Uwe Kühnberger**
({ahabdelfatta | tbesold | hgust | krumnack | martisch | kkuehnbe}@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück,
Albrechtstr. 28, 49076 Osnabrück, Germany

**Pei Wang**
(pei.wang@temple.edu)
Department of Computer and Information Sciences, College of Science & Technology, Temple University,
1805 N. Broad Street, Philadelphia, PA 19122 USA

## Abstract

The integration of artificial intelligence (AI) within cognitive science (CogSci) necessitates further elaborations on, and modelings of, several indispensable cognitive criteria. We approach this issue by emphasizing the close relation between artificial general intelligence (AGI) and CogSci, and discussing, particularly, "rationality" as one of such indispensable criteria. We give arguments evincing that normative models of human-like rationality are vital in AGI systems, where the treatment of deviations from traditional rationality models is also necessary. After conceptually addressing our rationality-guided approach, two case-study systems, NARS and HDTP, are discussed, explaining how the allegedly "irrational" behaviors can be treated within the respective frameworks.

**Keywords:** Rationality; intelligence; AGI; HDTP; NARS

## Motivations and Background

For more than five decades, artificial intelligence (AI) has always been a promising field of research on modeling human intelligence. The success of projects like IBM's Watson (Ferrucci et al., 2010), for instance, increases the hopes in achieving not only language intelligence but also inference mechanisms at a human-level and paves the way for solving more baffling tasks. However, AI has turned into a vague, unspecific term, in particular because of the tremendous number of applications that belong, in fact, to seemingly orthogonal directions. Philosophers, psychologists, anthropologists, computer scientists, linguists or even science fiction writers have disparate ideas as to what AI is (or should be). The challenge becomes more obvious when AI is looked at from a CogSci perspective, where the focus is mainly on explaining processes of general cognitive mechanisms (not only on how one or another intelligence task can be solved by a computer). We think that from a CogSci perspective the kind of intelligence characterizing classical AI problems is not yet exhaustive enough. Solutions to most of the problems are not cognitively inspired: neither do they consider essential cognitive mechanisms (or general intelligence results) nor do they show the biological plausibility of the solutions.

*Artificial General Intelligence* (AGI) refers to a research direction that takes AI back to its original goals of confronting the more difficult issues of human-level intelligence as a whole. Current AGI research explores all available paths, including theoretical and experimental computer science, cognitive science, neuroscience, and innovative interdisciplinary methodologies (Baum, Hutter, & Kitzelmann, 2010). Here, we approach cognition in AGI systems by particularly promoting "rationality" as one of such indispensable criteria, and analyze some divergent, sometimes seemingly irrational, behaviors of humans.

In this article, our goal is twofold. We first concern ourselves with explicitly allocating ideas from AGI within CogSci. Second, we give a conceptual account on some principles in normative rationality-guided approaches. After explaining our approach at a general level, we explain how two cognitively inspired systems, namely NARS and HDTP, have the potential to handle (ir)rationality. We conclude by giving some remarks and future speculations.

### Why AGI?

In current AGI research, there are approaches following different paths, including those (1) inspired by the structure of human brain or the behavior of human mind, (2) driven by practical demands in problem solving, or (3) guided by *rational principles* in information processing. We are concerned with the latter approach, which has at least three essential advantages. One advantage of the rationality-guided approach, from an AGI perspective, is that it is less bound to exactly reproducing human faculties on a functional level. Another advantage is that it gives AI the possibility of being established in a way similar to other disciplines, where it can give a theoretical explanation to intelligence as a process that can be realized both in biological systems and computational devices. The third advantage of the rationality-guided approach is that it is not limited to a specific domain or problem.

## Rationality

The term *rationality* is used in a variety of ways in various disciplines. In CogSci, rationality usually refers to a way a cognitive agent deliberatively (and attentively) behaves in, according to a specific normative theory. The prototypical instance of cognitive agents that can show rational behavior is humans, who so far are also the ultimate exemplar of generally intelligent agents. When modeling intelligence, it is reasonable to initially take the remarkable abilities of humans into account with respect to rational behavior, but also their apparent deficiencies that show up in certain tasks.

Surprisingly little attention has been paid so far in AI towards a theory of rationality. A reason might be that the concept of rationality was too broad in order to be of interest to AI, where for a long time usually relatively specific cognitive abilities were modeled and heuristics were suggested. Moreover, an artificial cognitive agent is usually intended to reproduce rational behavior, not to act in seemingly irrational ways. Consequently, AI researchers are not interested in results of some *classical rationality puzzles*. Still, we think that a move towards integrating AGI in CogSci cannot ignore rationality issues, neither the remarkable abilities nor the originalities human subjects show in rationality tasks.

## Traditional Models of Rationality

Different models of rationality use significantly different methodologies. Clustering such models according to the underlying formalism usually results in at least the following four classes: (1) logic-based models (Evans, 2002), (2) probability-based models (Griffiths, Kemp, & Tenenbaum, 2008), (3) heuristic-based models (Gigerenzer, 2008), and (4) game-theoretically based models (Osborne & Rubinstein, 1994). Several of these models have been proposed for establishing a *normative theory of rationality*, normally by judging a belief as rational if it has been obtained by a formally correct application of the respective reasoning mechanism, given some background beliefs or knowledge (cf. e.g. also (Gust et al., 2011; Wang, 2011)). Therefore, such theories of rationality are not only intended to model "rational behavior" of humans, but to postdictively decide whether a particular belief, action, or behavior is rational or not. Nonetheless, although a conceptual clarification of rational belief and rational behavior is without any doubts desirable, it is strongly questionable whether the large number of different (quite often orthogonal) frameworks makes this task easier, or if the creation of a more unified approach wouldn't be recommendable. From our perspective, basic cognitive mechanisms seem to offer a basis for such an endeavor.

## Some Rationality Challenges and Puzzles

Although the models mentioned above have been proven to be quite successful in modeling certain aspects of intelligence, all four types of models have been challenged. For example, in the famous Wason selection task (Wason & Shapiro, 1971) human subjects fail at a seemingly simple logical task (cf. Table 1.a). Similarly, Tversky and Kahneman's Linda problem (Tversky & Kahneman, 1983) illustrates a striking violation of the rules of probability theory in a seemingly simple reasoning problem (cf. Table 1.b). Heuristic approaches to judgment and reasoning try to stay closer to the observed behavior and its deviation from rational standards (Gigerenzer, 2008), but they fail in having the formal transparency and clarity of logic-based or probability-based frameworks with regard to giving a rational explanation of behavior. Game-based frameworks can be questioned due to the various forms of optimality concepts in game-theory that can support different "rational behaviors" for one and the same situation.

In order to make such challenges of rationality theories more precise, we discuss some aspects of the famous Wason selection task and the Linda problem in more detail.

**Wason Selection Task** This task shows that a large majority of subjects are seemingly unable to evaluate the truth of a simple rule of the form *"if p then q"* (Wason & Shapiro, 1971). In the version depicted in Table 1.a, this rule is represented by: *"If on one side of the card there is a D, then on the other there is the number 3"*. According to classical logic, in order to assign a truth-value to this rule, subjects need to turn D and 7. What is interesting is the fact that a slight modification of the content of the rule to a setting more familiar from daily life, while keeping the structure of the problem isomorphic, makes subjects perform significantly better, as e.g. shown in (Cosmides & Tooby, 1993).

Table 1: a. A description of the Wason selection task. b. An abbreviated version of the Linda problem setting.

**a. Wason Selection Task (Wason & Shapiro, 1971):**

Every card which has a D on one side has a 3 on the other side (and knowledge that each card has a letter on one side and a number on the other side), together with four cards showing respectively D, K, 3, 7, hardly any individuals make the correct choice of cards to turn over (D and 7) in order to determine the truth of the sentence. This problem is called "selection task" and the conditional sentence is called "the rule".

**b. Linda Problem (Tversky & Kahneman, 1983):**

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

(F): Linda is active in the feminist movement.

(T): Linda is a bank teller.

(T&F): Linda is a bank teller and is active in the feminist movement.

**Linda Problem** With respect to the Linda problem (Tversky & Kahneman, 1983) it seems to be the case that subjects have problems to prevent the so-called *conjunction fallacy*: subjects are told a story specifying a particular profile about someone called Linda. Then, some statements about Linda are shown and subjects are asked to order them according to their probability (cf. Table 1.b). 85% of subjects decide to rank the statements "*Linda is a bank teller and is active in the feminist movement*" (T & F) as more probable than the statement "*Linda is a bank teller*" (T). This ranking conflicts with the laws of probability theory, because the probability of two events (T & F) is less than or at most equal to the probability of one of the events (e.g. (T)).

## Classical Resolution Strategies of Irrationality

Many strategies have been proposed to address the mentioned challenges, ranging from the use of non-classical logics to model subjects' behavior in the Wason selection task (Stenning & van Lambalgen, 2008), to considerations involving reasoning in semantic models instead of (syntactic) deductions (Johnson-Laird, 1988) in the case of the Wason selection task. With respect to the Linda problem it has been argued that pure probability theory is not appropriate for addressing the problem properly, but a foundation of the analysis of this problem in coherence theories would be necessary (Pfeifer, 2008). Another resolution strategy applicable to both puzzles is to question whether tasks were appropriately phrased in the respective experiments. In the Wason selection task the "if-then" rule presented in natural language is usually not equivalent to its interpretation in classical logic, and in the Linda puzzle the term "probable" can be interpreted differently by the subjects (Gigerenzer, 2005). In any case, although there are many proposals to address the challenges, there is no generally accepted rationality concept available yet. Moreover, specific frameworks can address specific challenges, but do not generalize to the breadth of the mentioned problems.

For a generally intelligent cognitive system a question that can be raised is: *which principles of rationality can be transferred to and modeled in AGI systems, in order to achieve intelligence on a human scale?* We will argue for models that link rationality to the ability of humans to establish analogical relations (continuing a line of reasoning started in (Besold et al., 2011)), and to the ability to adapt to the environment by making good use of previously obtained experiences.

## Non-Standard, CogSci-Based Approaches

The two examples discussed above definitely show that humans have sometimes problems to apply rules of classical logic correctly (at least in rather abstract and artificial situations), and to reason according to the Kolmogorov axioms of probability theory. Nonetheless, the most that can be concluded from the experiments is that human agents are neither classical deduction machines nor probability estimators, but perform their indisputable reasoning capabilities by other means, necessarily linked to their cognitive capacities.

**Resolving the Selection Task by Cognitive Mechanisms** As mentioned above, subjects perform better (in the sense of more according to the laws of classical logic) in the Wason selection task, if content-change makes the task easier to access for subjects. We think that the performance of subjects has a lot to do with the ability of subjects to establish appropriate analogies. Subjects perform badly in the classical version of the Wason selection task, probably because they fail to establish a correct analogy. Therefore, subjects fall back to other (less reliable) strategies to solve the problem. In a content-change version of the task the situation is different, because subjects can do what they would do in an everyday analogous situation. In short, the success or failure of managing the task is crucially dependent on the possibility to establish a meaningful analogy.

Another related resolution is to study the mode of the inference that should underly a normative theory of rationality. When a system has sufficient knowledge and resources (with respect to the problems to be solved), an axiomatic logic (such as classical logic) can be used, which treats the available knowledge as axioms, and derives theorems from them to solve a given problem. When the system has insufficient knowledge, it has no absolute truth to be used as axioms, so has to follow some "non-axiomatic" logic, whose premises and conclusions are all revisable by new evidence. In Wason's task, the expected results are the ones assuming an axiomatic system, while the actual results may be consistent with a non-axiomatic one. Therefore, the "mistake" here is mainly the misunderstanding between the psychologists who run the tests and the subjects who take the tests. In this artificially structured experiment, it is valid for the psychologists to assume sufficient knowledge and resources, therefore to expect the application of an axiomatic type of inference mechanism. Their mistake, however, is the failure to see the result as coming from another type of inference. On the side of subjects, since non-axiomatic reasoning is used more often in everyday life, most of them fail to understand the experiment setting as a testing of their capacity of using an axiomatic inference mechanism. This explains why many subjects admit their mistake afterwards, and do better in the content-change task (as soon as they realized that the expected way of reasoning is not their default one, they have less problem to adapt to follow it).

**Resolving the Linda Problem by Cognitive Mechanisms** Here, a natural explanation of subjects' behavior is that there is a lower degree of coherence of Linda's profile plus the statement "*Linda is a bank teller*" in comparison to the degree of coherence of Linda's profile plus the statement "*Linda is a bank teller and is active in the feminist movement*", as in the conjunctive statement, at least one conjunct of the statement fits quite well to Linda's profile. *Coherence* (Thagard, 2002) is a complicated concept that needs to be discussed in more detail (as does its connection to notions like the idea of representativeness proposed as an explanation for the Linda problem by Tversky and Kahneman themselves), but it can be mentioned that coherence is important for the successful establishment of an analogical relation, as well as for guiding adaptation of obtained knowledge and experiences. In order to make sense out of the task, subjects tend to rate statements with a higher probability where facts are arranged in a theory with a higher degree of coherence. Also, this can be thought of as a form of coherently adapting beliefs, which also depends heavily on subjects' experiences rather than on their knowledge of Kolmogorov axioms of probability theory.

## Modeling Rationality: Case Studies

Formal and computational models in CogSci can be roughly divided into two major types: *descriptive* and *normative*. A descriptive model explains how a system actually works, and its establishment is based on empirical data. A descriptive model's quality is evaluated according to its behavior's *similarity* to that of humans. A normative model, on the other hand, specifies how a system should work, and its establishment is based on certain general principles or postulates. Such a normative model's quality is evaluated according to its behavior's *coherence* with these basic assumptions. Though the two types of models are closely related, they are still built and evaluated differently (Wang, 2011). When building a model of rationality, a central issue is the selection of the assumptions on which the model is based, since all conclusions about the model are derived from, and justified against, these assumptions.

In the following, we give two examples for cognitively inspired systems: NARS and HDTP. Both stand in a certain tradition to classical cognitive architectures like the well-known models ACT-R (Anderson & Lebiere, 1998) and SOAR (Laird, Newell, & Rosenbloom, 1987), because they attempt to model cognition in breadth and not relative to highly specialized abilities. Nevertheless, because NARS and HDTP stand in a tradition of modeling the competence aspect of general intelligence, they attempt to integrate a bunch of different human-inspired reasoning abilities, and they try to integrate these abilities in uniform models, they also differ significantly from the mentioned classical cognitive architectures. We briefly introduce NARS and HDTP and discuss how they can account for "irrational" behaviors in tasks, such as the Selection Task and the Linda problem.

**AGI with Relative Rationality (NARS)** NARS (Non-Axiomatic Reasoning System) is an AGI system designed under the assumption that the system usually has insufficient knowledge and resources with respect to the problems to be solved, and must adapt to its environment. Therefore, the system realizes a "relative rationality", that is, the solutions are the best the system can get *under the current knowledge–resource restriction* (Wang, 2011). Since this system has been described in a book (Wang, 2006) and many papers (most of which are available at the the last author's website[1]), here we only briefly explain the treatment of the "Selection Task" and "Conjunction Fallacy" in NARS.

Since NARS has insufficient knowledge and resources, its beliefs are not "absolute truth" but summary of the system's experience. Especially, the *truth-value* of a statement measures its *evidential support*, and the evidence can be either *positive* or *negative*, depending on whether the evidence agrees with the statement. Concretely, for statement "*If on one side of the card there is a D, then on the other there is the number 3*", the D card always provides evidence (positive if the other side is 3, otherwise negative); the 3 card may provide positive evidence (if the other side is D); the 7 card may provide negative evidence (if the other side is D); the K card provides no evidence. To determine the truth-value of the statement, all cards except K should be checked, but due to insufficient resources, the system may fail to recognize all evidence. In this case, D is the easiest, while 7 the hardest. This result is consistent with the common responses of human beings. It is labeled as "irrational", because in classical logic the truth-value of a statement only depends on the existence of *negative* evidence, and whether there is *positive* evidence does not matter. Furthermore, classical logic does not consider resource restriction at all. For a detailed discussion on evidence and truth-value in NARS, see (Wang, 2009).

In NARS, the meaning of a concept, such as "Linda" or "feminist bank-teller", is determined by the available information about it, in terms of how it relates to other concepts, as far as the system knows. For a given concept, such information may be either *extensional* (indicating its instances or *special cases*) or *intensional* (indicating its properties or *general cases*). To decide the extent to which a concept, "Linda", is a special case of another one, "bank-teller" or "feminist bank-teller", the system will consider all available evidence. In this example, the most accessible evidence about all three concepts are *intensional* (i.e., about their properties), so the system reaches its conclusion by checking if Linda has the properties usually associated with "bank-teller" and "feminist bank-teller", respectively. Since according to the given information Linda has more common properties with "feminist bank-teller" than with "bank-teller", her "degree of membership" is higher to the former than to the latter. This is judged as a "fallacy" when probability theory is applied *extensionally* to this situation, so only the *base rates* matters, while the properties do not. For a detailed discussion on the categorization model in NARS, see (Wang & Hofstadter, 2006).

In summary, as soon as a normative model of rationality or intelligence makes more realistic assumptions, many "heuristics", "bias", and even "fallacies" follow from them. In the above examples, there are strong reasons for assuming that the truth-value of a statement should depend on both positive and negative evidence (rather than negative only), and the meaning of a concept should depend on both extensional and intensional relations (rather than extensional only). We believe these examples mainly show the limitations of traditional models (classical logic, probability theory), rather than human errors. The practice of NARS and similar systems shows that it is possible for a new normative model to explain and reproduce similar results in a unified way.

**Rationality Through Analogy (HDTP)** As a second case study, we want to sketch how *Heuristic-Driven Theory Projection* (HDTP), an analogy-engine, can be used to implement some crucial parts of our cognitively-based theory of rationality (for an expanded elaboration cf. e.g. (Besold et al., 2012)). HDTP is a framework for computing analogical relations between two domains that are axiomatized in many-sorted first-order logic (Schwering, Krumnack, Kühnberger, & Gust, 2009). It provides an explicit generalization of the two domains as a by-product of establishing an analogy. Such a generalization can be a base for concept creation by abstraction. HDTP proceeds in two phases: in the *mapping phase*, the source and target domains are compared to find structural commonalities, and a generalized description is created, which subsumes the matching parts of both domains. In the *transfer phase*, unmatched knowledge in the source domain

[1] At `http://www.cis.temple.edu/∼pwang/papers.html`.

is mapped to the target domain to establish new hypotheses. HDTP is therefore similar in spirit to the well-known Structure-Mapping Engine (SME) (Falkenhainer, Forbus, & Gentner, 1989), e.g. with respect to the mentioned mapping and transfer phases and the symbolic representation of domains. Nevertheless, HDTP also differs significantly from SME, e.g. with respect to the strong expressive power of the underlying domain theories (many-sorted first-order logic in HDTP vs. propositional logic in SME), the establishment of the analogy relation as a by-product of an abstraction, and the massive usage of heuristics differ from the ones used in SME.

HDTP implements a principle (by using heuristics) that maximizes the coverage of the involved domains (Schwering et al., 2009). Intuitively, this means that the sub-theory of the source (or the target) that can be generated by re-instantiating the generalization is maximized. The higher the coverage the better, because more support for the analogy is provided by the generalization. A further heuristics in HDTP, for which the motivation is to prevent arbitrary associations, is the minimization of substitution lengths in the analogical relation, i.e. the simpler the analogy the better (Gust, Kühnberger, & Schmid, 2006). There is a trade-off between high coverage and simplicity of substitutions: An appropriate analogy should intuitively be as simple as possible, but also as general and broad as necessary, in order to be non-trivial. This kind of trade-off is similar to the trade-off that is usually the topic of model selection in machine learning and statistics.

The modeling of the Wason selection task with HDTP is quite simple as long as appropriate background knowledge is available, in case an analogy should be established, or the lack of appropriate background knowledge prevents analogy making, in case no analogy should be established. In other words, the availability of appropriate resources in form of background knowledge is crucial. If appropriate background knowledge for an analogous case is missing, then there is no chance to establish an analogical relation or a potential analogy (with low coverage and complex substitutions) is misleading the subject. Hence, subjects have to apply other strategies. This is the situation when subjects are confronted with the original Wason selection task based on properties of cards. Most subjects have problems to establish a meaningful analogy with a well-known domain due to the high degree of abstractness of the task itself. In the other case, if there is a source theory with sufficient structural commonalities, then the establishment of an analogical relation is straightforward. This happens if the task is changed in the following way: the rule that needs to be checked is now: *"If someone is drinking beer in a bar, he / she must be older than 21"*. In the experiment, subjects can choose between "drinking beer", "drinking coke", "25 years old", and "16 years old" (Cosmides & Tooby, 1993). In the corresponding experiments, subjects behave significantly better than in the original selection task. With analogy making the improvement of the subjects in mastering the task can be explained. They can establish an analogy between the sketched set-up of the experiment and a standard situation in daily life, in which they would simply do the necessary actions to check whether there is someone who is drinking beer in the bar without being older than 21: check people who are drinking beer, and check what people are drinking who are 16. As both situations are very similar to each other, the generalization is straightforward, substitutions length are minimal, and coverage is high.

The Linda problem is structurally different in comparison to the Wason selection task. In an analogy making context, an explanation of subjects' behavior in terms of coherence maximization is promising. Coherence aspects of input theories are crucial for establishing analogies in several ways. Roughly speaking, the statement *"Linda is a bank teller"* has less coherence with Linda's profile than the statement *"Linda is a bank teller and is active in the feminist movement"*. Therefore, it is easier to establish an analogy between Linda as given in Linda's profile and Linda as described in *"Linda is a bank teller and is active in the feminist movement"* than in the pure "bank teller" case. Notice that from an abstract point of view the coherence-based resolution of the task is rather similar with the intensional interpretation of the task in NARS, where "feminist bank teller" has a higher degree of membership with Linda's profile than "bank teller".

## Conclusion and Future Work

There are multiple models of rationality, each with its own assumptions and applicable situations. The traditional models are based on certain idealized assumptions, and thus are limited to the domains where the latter are satisfied. Since human cognition has evolved in and is usually used in realistic situations where those idealized assumptions do not hold, those models of rationality are not universally applicable, and violations should not be deemed "irrational" per se. The seemingly irrational behaviors are there not because the intelligent systems (e.g. humans) are irrational, but because the traditional normative theories do not cover rationality very well.

Instead, we believe what is needed are new models of rationality that are based on more realistic assumptions and developed in a more holistic framework. Such models should be able to provide an adequate and feasible positive account of actual human rationality, also accommodating particularities of human-style reasoning. Such a framework could form a cornerstone of a closer connection between AGI and CogSci, embedding important parts of the AGI program within a CogSci context, whilst making the more general methods and theories of AGI accessible to the CogSci side.

The overall appeal for a "more cognitive" view on rationality models and systems is infrequent, but not unusual. Amongst others, already Kokinov (2003) reaches the conclusion that the concept of rationality as a theory in its own right ought to be replaced by a multilevel theory based on cognitive processes involved in decision-making. On the more technical side, there is a growing body of evidence that analogy engines (like HDTP) and general-purpose reasoning engines (like NARS) can be used for implementing these cognitive mechanisms and, thus, also as foundations of a rationality-guided approach to general intelligence.

This paper should merely be considered as a point of departure, leaving questions for future research galore. For example with respect to the present proposal concerning HDTP, it seems recommendable to figure out to which extent different types of coherence concepts can be integrated into the framework. In particular, the challenges mentioned above need to be addressed, and a formal treatment of coherence needs to be fleshed out. Furthermore, an implementation of coherence principles for retrieval, mapping, and re-representation purposed in the analogy making process needs to be formulated. Concerning NARS, amongst others the following issues would merit work and effort: real-time temporal inference, procedural inference, and self-control. Regarding competing theories for rationality, clarifying to what extent cognitive capacities and limitations have already been taken into account (implicitly as well as explicitly) when designing the theories, and to what extent the classical frameworks can be re-instantiated by a cognitively-based approach, has to be considered one of the principal questions for future research. Finally, on a fundamental conceptual level, a broader definition of rational beliefs is still needed.

## References

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Baum, E., Hutter, M., & Kitzelmann, E. (Eds.). (2010). *Artificial General Intelligence*. Lugano, Switzerland: Atlantis Press.

Besold, T. R., Gust, H., Krumnack, U., Abdel-Fattah, A., Schmidt, M., & Kühnberger, K. (2011, July). An Argument for an Analogical Perspective on Rationality & Decision-Making. In J. van Eijck & R. Verbrugge (Eds.), *Proc. of the Workshop Reasoning About Other Minds (RAOM-2011).* CEUR-WS.org, Vol. 751.

Besold, T. R., Gust, H., Krumnack, U., Schmidt, M., Abdel-Fattah, A., & Kühnberger, K.-U. (2012). Rationality Through Analogy - Towards a Positive Theory and Implementation of Human-Style Rationality. In I. Troch & F. Breitenecker (Eds.), *Proc. of MATHMOD 12 Vienna.*

Cosmides, L., & Tooby, J. (1993). Cognitive adaptations for social exchange. In J. H. Barkow and L. Cosmides and J. Tooby (Ed.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163–228). Oxford.

Evans, J. (2002). Logic and Human Reasoning: An Assessment of the Deduction Paradigm. *Psychological Bulletin*, *128*, 978–996.

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and Example. *Artificial Intelligence*, *41*, 1–63.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., et al. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, *31*(3), 59–79.

Gigerenzer, G. (2005). I think, therefore I err. *Social Research*, *72*(1), 195–218.

Gigerenzer, G. (2008). *Rationality for Mortals: How People Cope with Uncertainty*. Oxford University Press.

Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian Models of Cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Cognitive Modeling.* Cambridge University Press.

Gust, H., Krumnack, U., Martínez, M., Abdel-Fattah, A., Schmidt, M., & Kühnberger, K.-U. (2011). Rationality and General Intelligence. In J. Schmidhuber, K. Thorisson, & M. Looks (Eds.), *Artificial General Intelligence* (pp. 174–183).

Gust, H., Kühnberger, K.-U., & Schmid, U. (2006). Metaphors and Heuristic-Driven Theory Projection (HDTP). *Theor. Comput. Sci.*, *354*, 98–117.

Johnson-Laird, P. (1988). *Cognitive science*. Cambridge University Press.

Kokinov, B. (2003). Analogy in Decision-Making, Social Interaction, and Emergent Rationality. *Behavioral and Brain Sciences*, *26*(2), 167–168.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An Architecture for General Intelligence. *Artificial Intelligence*, 1–64.

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

Pfeifer, N. (2008). A Probability Logical Interpretation of Fallacies. In G. Kreuzbauer, N. Gratzl, & E. Hiebl (Eds.), *Rhetorische Wissenschaft: Rede und Argumentation in Theorie und Praxis* (pp. 225–244). LIT-Verlag.

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

Stenning, K., & van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. MIT Press.

Thagard, P. (2002). *Coherence in Thought and Action*. MIT Press.

Tversky, A., & Kahneman, D. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, *90*(4), 293–315.

Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence*. Dordrecht: Springer.

Wang, P. (2009). Formalization of Evidence: A Comparative Study. *Journal of Artificial General Intelligence*, *1*, 25–53.

Wang, P. (2011). The Assumption on Knowledge and Resources in Models of Rationality. *International Journal of Machine Consciousness (IJMC)*, *3*, 193–218.

Wang, P., & Hofstadter, D. (2006). A Logic of Categorization. *Journal of Experimental & Theoretical Artificial Intelligence*, *18*(2), 193–213.

Wason, P. C., & Shapiro, D. (1971). Natural and Contrived Experience in a Reasoning Problem. *Quarterly Journal of Experimental Psychology*, *23*, 63–71.