Rationality and General Intelligence

Helmar Gust, Ulf Krumnack, Maricarmen Martínez, Ahmed Abdel-Fattah, Martin Schmidt, and Kai-Uwe Kühnberger

University of Osnabrück, Albrechtstr. 28, Germany
{hgust,krumnack,mmartine,ahabdelfatta,martisch,kkuehnbe}@uos.de

Abstract. Humans are without any doubts the prototypical example of agents that can hold rational beliefs and can show rational behavior. If an AGI system is intended to model the full breadth of human-level intelligence, it is reasonable to take the remarkable abilities of humans into account with respect to rational behavior, but also the apparent deficiencies of humans in certain rationality tasks. Based on well-known challenges for human rationality (Wason-Selection task and Tversky & Kahneman's Linda problem) we propose that rational belief of humans is based on cognitive mechanisms like analogy making and coherence maximization of the background theory. The analogy making framework *Heuristic-Driven Theory Projection* (HDTP) can be used for implementing these cognitive mechanisms.

Keywords: Rationality, Analogy, Coherence, HDTP.

1 Introduction

Although human behavior can seem erratic and irrational at times, only few people would doubt that human behavior can be rational and, in fact, appears rational most of the time. If we explain behavior, we use terms like beliefs and desires. If an agent's behavior makes the most sense to us, then we interpret it as a reasonable way to achieve the agent's goals given his beliefs. Hence, the concept of rationality and, in particular, the epistemic aspects of rationality, namely the consideration of rational beliefs of an agent, does play a crucial role in describing and explaining behaviors of humans in a large variety of situations.

Discussions about and theories on rationality are often linked to disciplines like psychology, economy, and philosophy. Little attention has been paid so far in artificial (general) intelligence towards a theory of rationality, although a currently increasing endeavor in AI and AGI to model generalized forms of intelligence cannot be denied.[1] A reason might be that the concept of rationality is too broad in order to be of interest to artificial intelligence, where usually relatively specific cognitive abilities are modeled. Another reason might be the

[1] Cf. [11] or [22] for two examples intended to model general intelligence. Major differences between rationality issues as discussed in this paper and models for general intelligence are the focus on cognitive mechanisms and the inspiration of seemingly irrational behavior of human subjects in the current proposal.

J. Schmidhuber, K.R. Thórisson, and M. Looks (Eds.): AGI 2011, LNAI 6830, pp. 174–183, 2011.

lack of interest of AI researchers concerning classical rationality puzzles, because an artificial agent is intended to reproduce rational behavior, but is not intended to reproduce seemingly irrational human behavior (cf. Section 2 for a discussion of some of these puzzles). Nevertheless, we think that a move towards artificial *general* intelligence cannot ignore any longer rationality issues of human subjects. In particular, this means that neither the remarkable abilities nor the obvious deficits human subjects show in rationality tasks should be ignored. For a general intelligent system the question that can be raised is which properties of rationality can be transferred to and modeled in AGI frameworks, in order to achieve intelligence on a human scale.

Although, even in psychology or economics there is no generally accepted formal framework for rationality, we will try to argue for a model that links rationality to the ability of humans to establish analogical relations. This is an attempt for proposing a new perspective and framework for rationality. Our argumentation is mostly conceptual in nature and not empirically based. Nevertheless, we think that there are strong conceptual arguments for linking rationality and analogy making.

2 Rationality Concepts and Challenges

2.1 Rationality

Many frameworks have been proposed for modeling rationality. Different frameworks for rationality use significantly different methodologies. Clustering such frameworks results in at least the following four classes.

- Logic-based models (cf. e.g. [3])
- Probability-based models (cf. e.g. [7])
- Heuristic-based models (cf. e.g. [6])
- Game-theoretically based models (cf. e.g. [15])

Several of these models have been proposed for establishing a normative theory of rationality. For example, with respect to logic theories, this means in its simplest form that a belief is rational, if there is a logically valid reasoning process to reach this belief (relative to available and given background knowledge). With respect to probabilistic approaches, a belief is rational, if the expectation value of this belief is maximized (relative to given probability distributions of background beliefs). Therefore, such theories of rationality are not only intended to model "rational behavior" of humans, but also to predictively determine whether a particular belief, action, or behavior is rational or not. Furthermore, such theories specify *definitions* of rationality. Although a conceptual clarification of rational belief and rational behavior is without any doubts desirable, it is questionable whether the large number of different (and quite often orthogonal) frameworks make this task easier. In this paper, we will not try to propose a new (normative) definition of rational belief. Rather, we propose to explain and specify rationality and rational belief of human subjects by referring to certain cognitive mechanisms like analogy making and coherence maximization of the background theory. Furthermore, we intend to show that such mechanisms can be implemented and modeled computationally.

Table 1. The Wason-selection task questions whether humans reason in such situations according to the laws of classical logic. Tversky and Kahneman's Linda problem questions the ability of humans to reason according to the laws of probability theory.

Wason-Selection Task [24]:
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 the "selection task" and the conditional sentence is called "the rule".
Linda-Problem [21]:
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. Linda is a teacher in elementary school. Linda works in a bookstore and takes Yoga classes. Linda is active in the feminist movement. (F) Linda is a psychiatric social worker. Linda is a member of the League of Women Voters. Linda is a bank teller. (T) Linda is an insurance salesperson. Linda is a bank teller and is active in the feminist movement. (T&F)

2.2 Well-Known Challenges

Although the classes of frameworks mentioned in Section 2.1 have been proven to be quite successful in modeling certain aspects of human intelligence, they have been challenged by psychological experiments. For example, in the famous Wason-selection task [23] human subjects fail at a seemingly simple logical task (cf. Table 1). Similarly, Tversky and Kahneman's Linda problem [21] illustrates a striking violation of the rules of probability theory in a seemingly simple reasoning problem (cf. Table 1). Heuristic approaches to judgment and reasoning [5] try to stay closer to the observed behavior and its deviation from rational standards. They are often seen as approximations to a rational ideal or at least sometimes can be demonstrated to work in practice, but they fail in having the same formal transparency and clarity of logic-based or probability-based frameworks with regard to giving a rational explanation of behavior. Game-based frameworks are questioned due to the various forms of optimality concepts in game-theory that can support different "rational behaviors" for one and the same situations (e.g. Pareto optimality vs. Nash equilibrium vs. Hick's optimality [1]).

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 verify or to falsify a simple logical rule of the form "$p \rightarrow q$". In the version depicted in Table 1, 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". In order to check this rule, subjects need to turn D and 7, i.e. subjects need to check the direct rule application and the contrapositive implication (*modus tolens* of the rule). What is interesting is the fact that a slight modification of the content of the rule (content-change), while keeping the structure of the problem isomorphic, subjects perform significantly better: In [2], the authors show that a slight change of content of the abstract rule "$p \rightarrow q$" to a well-known problem shows different results with a significant increase of correct answers of subjects. The authors use the rule "If a person is drinking beer, then he must be over 20 years old." The cards used in the task were "drinking beer", "drinking coke", "25 years old", and "16 years old". Solving this task according to the rules of classical logic comes down to turning "drinking beer" and "16 years old".

Linda Problem. With respect to the Linda problem 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 the bank teller Linda. Then, eight statements about Linda are shown and subjects are asked to order them according to their probability (cf. Table 1). 85% of subjects decide to rank the eighth statements "Linda is a bank teller and active in the feminist movement" (T & F) as more probable than the sixth statement "Linda is a bank teller" (T). This ranking is conflicting with the laws of probability theory, because the probability of two events (T & F) is less or at most equal to the probability of one of the events (e.g. T).

Classical Resolution Strategies. Many strategies have been proposed to address the mentioned challenges. For example, logicians proposed non-classical logics to model subjects' behavior in the Wason-Selection task [18]. Other researchers claim with respect to the Wason-Selection task that humans do not perform (syntactic) deductions, but do perform reasoning in semantic models [12]. For other challenges, like the Linda problem, again many strategies towards a solution have been proposed. Nevertheless, there is no generally accepted rationality concept available, yet. Moreover, specific frameworks can address specific challenges, but do not generalize in order to address the breadth of the mentioned problems. This situation is not very satisfying.

2.3 Non-standard Interpretations of Wason and the Linda Problem

An immediate reaction to the two mentioned challenges for rationality depicted above may be to deny that humans are able to correctly reason according to the laws of classical logic or the laws of probability theory. Nevertheless, we think that humans are remarkably smart. The two cases definitely show that humans have sometimes problems to apply rules of classical logic correctly (at least in rather abstract and artificial situations) and it also shows that they have

sometimes problem to reason according to the Kolmogorov axioms of probability theory. Whether this means that their behavior is therefore irrational is not so clear. The most that can be concluded from the experiments is that human agents are neither deduction machines nor probability estimators, but perform their undisputable reasoning capabilities with other means. Moreover, we think that the deeper reason for subjects' behavior in the described tasks is connected to certain cognitive mechanisms that are used by humans in such reasoning tasks.

Resolving the Wason-Selection Task by Cognitive Mechanisms. As mentioned above, according to [2] 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 have 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, simply because they fail to establish the right analogy. Therefore, subjects fall back to other strategies to solve the problem. In the "beer drinking" version mentioned above [2], i.e. the content-change version of the task, the situation is different, because subjects can do what they would do in an everyday analogous situation: they need to check whether someone younger than 20 years is drinking beer in a bar. This is to check the age of someone who is drinking beer and conversely to check someone who is younger that 20 years whether he is drinking beer or not. In short, the success or failure of managing the task is crucially dependent on the possibility to establish a meaningful analogy.

Resolving the Linda Problem by Cognitive Mechanisms. In case of Tversky and Kahneman's Linda problem, 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 coherence of Linda's profile plus the statement "Linda is a bank teller and is active in the feminist movement". In the conjunctive statement, at least one conjunct of the statement fits quite well to Linda's profile. In short, subjects prefer situations that seem to have a stronger inner coherence. Coherence is a complicated concept that will be discussed below in more detail, but it can be mentioned already that coherence is important for the successful establishment of an analogical relation. 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.

3 Rationality and AGI Systems

3.1 Modeling Rationality

One could object that an AGI system that attempts to implement rationality should not be based on mechanisms that seemingly trigger irrational beliefs like the ones shown in the Wason-Selection task. Nevertheless, this does not take into account that mechanisms like the ability to establish analogical relations or

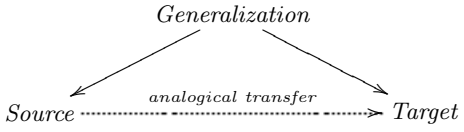

Fig. 1. HDTP's overall approach to creating analogies

the ability to maximize coherence properties of background theories can be seen as the very reason for many remarkable cognitive achievements of humans. We just mention some aspects with respect to analogy making:

- The ability to establish analogical relations can be interpreted as the reason "why we're so smart" [4].
- Analogy making is an important aspect of reasoning and "a core of cognition" [9].
- Analogy making can be taken as a framework for creativity [10].
- Analogy making is important for concept learning [14].

We think that intelligence on a human scale can only be reached, if such mechanisms like analogy making and maximizing coherence of theories are carefully taken into account. As a side-effect they can be used to explain some seemingly irrational behavior and decisions of subjects in tasks like the ones mentioned above. It should be mentioned that we do not claim that rational beliefs of natural and artificial agents are exclusively based on these two mechanisms. Other mechanisms like classical forms of reasoning (e.g. deduction and abduction), concept blending, reinforcement learning etc. are also necessary to reach a complete picture of cognition. Nevertheless, we claim that for the particular issues of rationality discussed in this paper, both mechanisms are crucial.

In order to give some hints how an analogy engine implements such cognitive mechanisms we sketch some basic ideas of *Heuristic-Driven Theory Projection* (HDTP) as an example of a powerful analogy making system. HDTP is a framework for computing analogical relations between two domains that are axiomatized in first order logic [17]. HDTP 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 can be mapped to the target domain to establish new hypotheses, cf. Figure 1.

HDTP implements a principle (by using heuristics) that maximizes the coverage of the involved domains [17]. Intuitively, this means that the sub-theory of the source (or the target) that can be generated by re-instantiating the generalization is maximized (cf. Figure 1). The higher the coverage the better, because more support for the analogy is provided by the generalization. A further heuristics in

HDTP is the minimization of substitution lengths in the analogical relation, i.e. the simpler the analogy the better [8]. The motivation for this heuristics is to prevent arbitrary associations. Clearly there is a trade-off between high coverage and simplicity of substitutions, or to put it differently, 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 kind of trade-off that is usually the topic of model selection in machine learning and statistics.

HDTP has been applied to a variety of domains [17], [9]. A 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: if background knowledge for an analogous case is missing, then there is no chance to establish an analogical relation, hence subjects have to apply other strategies. If there is a source theory with sufficient structural commonalities, then the establishment of an analogical relation is straightforward.

The Linda problem is structurally different in comparison to the Wason-Selection task. We think that an explanation of subjects' behavior in terms of coherence maximization, as sketched in the next subsection, is promising.

3.2 Coherence and Analogies

Basic concepts of coherence are discussed in several papers by Paul Thagard, cf. [19], [20]. In Thagard's approach, coherence is a property of sets of propositions (pieces of information) that is induced by the coherence values between two single propositions. Principles of coherence are formulated as a multi-constraint network of highly interconnected elements. The nodes of the network are pieces of information (e.g. formulas of a theory) and the (undirected) edges are weighted with coherence values. Positive values between two propositions support the coexistence of these pieces of information in the same theory, thereby increasing the global coherence of such a network, while negative values enforce decisions between alternatives (accepting only one of the items as part of the theory) or decrease the global coherence of the network if both pieces are included in the same theory. Hence, maximizing coherence means putting together those pieces of information that have a positive values between them while separating those having negative values.

In [20], the coherence values must fulfill certain constraints. The author gives four general constraints that are important for all types of coherence:

- The coherence between two propositions is symmetric.
- The coherence between contradictory items is negative.
- The acceptance of a proposition depends on the change in coherence if it is added.
- Propositions that are intuitively obvious have a degree of acceptability on their own.

Thagard proposed several types of coherence, for example, deductive, explanatory, conceptual, analogical, visual etc. coherence. Depending on the particular type of coherence additional constraints are proposed. Due to space limitations we cannot introduce the details of these concepts. Just to give the reader a flavor of the approach, we mention the constraints for deductive coherence:

- A proposition coheres with propositions that are deducible from it.
- Propositions that together are used to deduce some other proposition cohere with each other.
- The more hypotheses it takes to deduce something, the lower the coherence between them.

We think that there are three possibilities to support the analogical reasoning process by taking into account coherence.

- The maximization of coherence can be fruitfully used in order to extract a source domain for analogy making. This means that relevant entries of the underlying knowledge base need to be identified, selected, and retrieved.
- The mapping process incrementally builds the generalization of the underlying input theories. Maximizing the coherence of this generalization can be used as a control strategy for the mapping phase.
- With respect to the control of the transfer phase coherence of the target domain can indicate when to stop adding new formulas to the target.

Finally, we want to address the (open) question how coherence of theories is related to the two guiding principles used in HDTP, namely to maximize coverage and to minimize the complexity of analogical relations (i.e. minimize substitution lengths). The link between deductive coherence and the two HDTP principles is not straightforward, because there are obvious tensions between Thagard's constraints on coherence and the principles used in HDTP. Three challenges are mentioned in the following:

1. Coherence can be defined on finite or infinite sets of formulas, whereas the original coverage concept of HDTP is operating only on infinite theories, i.e. the deductive closure of an axiom system.
2. Coherence of sets of formulas is symmetric in Thagard's approach, whereas analogical relations are commonly considered to be non-symmetric.
3. Analogical associations are broader than coherence relations, because they can be productive, resulting in the creation of new concepts on the target.

We have currently no ready-made solutions for these challenges, but we add some speculations about possible answers. Challenge 1. can be addressed by the introduction of a modified notion of "finite coverage" for the HDTP framework. Naturally, this finite version of coverage would correspond to the re-represented inputs of the domain theories triggered by the analogical alignment. In order to address challenge 2. a careful assessment of the symmetry constraint in Thagard's approach and the commonly assumed non-symmetry of the alignment process

in analogy making needs to be carried out. It is relatively clear from research on analogy that the directedness of an analogical relation can be relaxed in certain circumstances. In particular, with respect to the creation of new concepts on the target domain (e.g. in cross-domain analogies) an adaptation between source and target is necessary specifying the parameters in which an analogy is appropriate. Hence, there is a reciprocal relation between source and target without a strict directedness of the analogical relation. Finally, challenge 3. requires an extended definition of coherence because analogies allow for creative transfers and the introduction of new concepts. Simple measures of coherence that are defined for fixed sets of propositions are therefore not suitable.

Conceptually, it is rather clear that every analogical relation between a source and a target domain is strongly dependent on a high coherence of the input theories as well as the coherence that is established by the analogy itself. We think that an integrated approach of both aspects in one framework is plausible.

4 Conclusions

In this paper, we proposed to introduce new aspects of rationality into the AGI context. Rationality plays an important role in different scientific disciplines, but did not get sufficient attention in AI or AGI. Based on the proposed new resolution strategies for classical rationality puzzles, we think that the usage of analogy making frameworks and theories for maximizing the coherence of a theory are good candidates for the implementation of rational belief. Although coherence theories in the tradition of Thagard and analogy making frameworks may seemingly be quite different frameworks, we claim that it is possible to instantiate a high degree of coherence of a theory in an analogy making framework.

We think that this paper is just a first conceptual step towards a theory of artificial rational agents. With respect to the present proposal, it is necessary to figure out to which extent different types of coherence concepts can be integrated into the HDTP framework. In particular, the challenges mentioned in Section 3.2 need to be addressed. A formal treatment of coherence in HDTP needs to fleshed out. Furthermore, an implementation of coherence principles for retrieval, mapping, and re-representation purposes in the analogy making process needs to be formulated. With respect to competing theories for rationality, it would be desirable to have formal approaches for heuristic approaches or game-theoretic approaches as well.

References

1. Chinchuluun, A., Pardalos, P.M., Migdalas, A., Pitsoulis, L. (eds.): Pareto Optimality, Game Theory and Equilibria. Springer, New York (2008)
2. Cosmides, L., Tooby, J.: In: Barkow, et al. (eds.) Cognitive Adaptations for Social Exchange. Oxford University Press, New York (1992)
3. Evans, J.S.B.T.: Logic and human reasoning: An assessment of the deduction paradigm. Psychological Bulletin 128, 978–996 (2002)

4. Gentner, D.: Why we're so smart. In: Goldin-Meadow, S. (ed.) Language in mind: Advances in the study of language and thought, pp. 195–235. MIT Press, Cambridge (2003)
5. Gigerenzer, G.: Rationality for Mortals: How People Cope with Uncertainty. Oxford University Press, Oxford (2008)
6. Gigerenzer, G., Hertwig, R., Pachur, T. (eds.): Heuristics: The Foundation of Adaptive Behavior. Oxford University Press, New York (in press)
7. Griffiths, T., Kemp, C., Tenenbaum, J.: Bayesian Models of Cognition. In: Sun, R. (ed.) The Cambridge Handbook of Computational Cognitive Modeling. Cambridge Univeresity Press, Cambridge (2008)
8. Gust, H., Kühnberger, K.-U., Schmid, U.: Metaphors and Heuristic-Driven Theory Projection (HDTP). Theoretical Computer Science 354, 98–117 (2006)
9. Gust, H., Krumnack, U., Kühnberger, K.-U., Schwering, A.: Analogical Reasoning: A Core of Cognition. Künstliche Intelligenz 1(08), 8–12 (2008)
10. Hofstadter, D., and the Fluid Analogies Research Group.: Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought. Basic Books, Inc., New York (1995)
11. Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Heidelberg (2005)
12. Johnson-Laird, P.: Mental Models. Harvard University Press, Cambridge (1983)
13. Krumnack, U., Schwering, A., Gust, H., Kühnberger, K.-U.: Restricted Higher-Order Anti-Unification for Analogy Making. In: Orgun, M.A., Thornton, J. (eds.) AI 2007. LNCS (LNAI), vol. 4830, pp. 273–282. Springer, Heidelberg (2007)
14. McLure, M., Friedman, S., Forbus, K.: Learning concepts from sketches via analogical generalization and near-misses. In: Proceedings of the 32nd Annual Conference of the Cognitive Science Society (CogSci), Portland, OR (2010)
15. Osborne, M., Rubinstein, A.: A Course in Game Theory. MIT Press, Cambridge (1994)
16. Schwering, A., Krumnack, U., Kühnberger, K.-U., Gust, H.: Analogy as Integrating Framework for Human-Level Reasoning. In: Wang, P., Goertzel, B., Franklin, S. (eds.) Artificial General Intelligence: Proceedings of the First AGI Conference, pp. 419–423. IOS, Memphis (2008)
17. Schwering, A., Krumnack, U., Kühnberger, K.-U., Gust, H.: Syntactic Principles of Heuristic-Driven Theory Projection. Cognitive Systems Research 10(3), 251–269 (2009)
18. Stenning, K., van Lambalgen, M.: Human Reasoning and Cognitive Science. MIT Press, Cambridge (2008)
19. Thagard, P.: Explanatory Coherence. Behavioral and Brain Sciences 12(3), 435–467 (1989)
20. Thagard, P.: Coherence in Thought and Action. MIT Press, Cambridge (2002)
21. Tversky, A., Kahneman, D.: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review 90(4), 293–315 (1983)
22. Wang, P.: Rigid Flexibility: The Logic of Intelligence. Springer, Heidelberg (2006)
23. Wason, P.C.: Reasoning. In: Foss, B. (ed.) New Horizons in psychology, Penguin, Harmondsworth (1966)
24. Wason, P.C., Shapiro, D.: Natural and contrived experience in a reasoning problem. The Quarterly Journal of Experimental Psychology 23(1), 63–71 (1971)