Chapter 3
Rationality {in|for|through} AI

Tarek 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.

3.1 Introduction

Human behavior sometimes seems erratic and irrational. Nonetheless, it is widely undoubted that man can act rational and actually appears to act rational most of the time, making him an *animal rationale*. In explaining behavior, often terms like beliefs and desires are used: 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. This can be taken as indication that some concept of rationality does play a crucial role when describing and explaining human behavior in a large variety of situations.

Now, being faithful to the original goal of AI, namely the (re)creation of an artificially intelligent agent being at least comparable to humans in terms of intelligence and related notions, the step to considering rationality an important concept also within the context of research in artificial intelligence is a small one. If a machine should be considered as at least human-like in its capabilities and behavior, a human agent would most likely be taken as comparandum, and the alleged (ir)rationality of the machine's actions and behavior would be judged in comparison to the corresponding human examples.

Tarek R. Besold
Institute of Cognitive Science, University of Osnabrück, 49069 Osnabrück, Germany
e-mail: tbesold@uni-osnabrueck.de

J. Kelemen et al. (Eds.): Beyond Artificial Intelligence, TIEI 4, pp. 49–62.
DOI: 10.1007/978-3-642-34422-0_3

Therefore, in the following, I will investigate into and reflect on the role rationality has played in the history of AI this far, will provide an assessment of the actual status quo, and will also give some proposals and recommendations for a possible future of rationality and corresponding research in relation to AI.

Sect. 3.2 offers a general introduction to previous work and theories of rationality in general, naming the four classical paradigms for modeling rationality and rational behavior, before Sect. 3.3 elaborates on rationality within the context of AI in more detail, amongst others providing motivations for work on rationality in AI, as well as an assessment of the actual situation and status quo. Sect. 3.4 then sketches the foundations of a proper research program on theories and models of rationality within AI, providing arguments in favor of such an endeavor, before Sect. 3.5 presents a vision of how foundations for a model of rationality suitable for the aims and purposes of AI might look like. Sect. 3.6 summarizes some of the main aspects of the proposed stance in a short conclusion.

3.2 Rationality

Research in rationality and rational behavior has a rich history, both within more abstract disciplines like philosophy (possibly starting with Aristotle's ascription of a rational principle to the human being in his Nicomachean Ethics [6]) or economics (for instance think of the model of the *homo oeconomicus*, conceptually at least dating back to a 1836 essay by John Stuart Mill [25]), as well as in more individual-centered fields as psychology and cognitive science (several examples are given in the following).

Over the years, many quite distinct frameworks for modeling rationality (and establishing a normative theory) have been proposed. Breaking these distinct approaches down to their underlying theoretical foundations, four main types of models can be identified, together with corresponding normative interpretations for what has to be judged rational according to the respective approach:[1]

- Logic-based models (cf. e.g. [13]): A belief is rational, if there is a logically valid reasoning process to reach this belief relative to available/given background knowledge.
- Probability-based models (cf. e.g. [20]): A belief is rational, if the expectation value of this belief is maximized relative to given probability distributions of background beliefs.
- Game-theoretically based models (cf. e.g. [26]): A belief is rational, if the expected payoff of maintaining the belief is maximized relative to other possible beliefs.
- Heuristic-based models (cf. e.g. [17]).

[1] Partially with exception of the heuristic-based models, as due to their fundamentally different approach to conceptualizing and understanding rationality classical ideas of normativity within a model of rationality become obsolete and in many cases are not anymore part of the respective frameworks.

Unfortunately, it shows that the definitions of rationality arising from within the distinct approaches are in many cases almost orthogonal to each other (as are the frameworks), making them in the best case incommensurable, if not inconsistent or even partly contradictory in their modeling assumptions. Also, in many cases the predictive power of these classical theories of rationality turns out to be rather limited (at least when applied to real-world examples instead of artificially simplified and constructed scenarios), as their emphasis clearly lies on the normative or postdictive-explanatory aspects of the respective models – which should be a crucial observation when thinking about rationality and its role and applications in AI.

Even more, although each of the listed accounts has gained merit in modeling certain aspects of human rationality, the generality of each such class of frameworks has at the same time been challenged by psychological experiments or theoretical objections:

- On the one hand, studies by Wason and Shapiro question the human ability of reasoning in accordance with the principles of classical logic [38], and also Byrne's findings on human reasoning with conditionals [8] indicate severe deviations from this classical paradigm.

  Similarly, when considering probability-based models, Tversky and Kahneman's Linda problem [36] illustrates a striking violation of the rules of probability theory.
- On the other hand, game-based frameworks are questionable due to the lack of a unique concept of optimality in game-theory. Provided e.g. with the plethora of different equilibrium concepts which have been derived from the original Nash equilibrium (cf. e.g. [21]), which one shall be taken as "'the most rational one" given a certain situation?
- Finally, heuristic approaches to judgment and reasoning follow a different approach, and are often conceptualized as approximations to a rational ideal, rather than an instantiation of the ideal itself. In some cases and application scenarios, heuristics actually do work well and can yield surprising results in practice, but still mostly lack formal transparency and explanatory power. Moreover, their status as almost magical solutions to problems relating with complexity and perceived intractability in human decision-making has recently been challenged in a quite fundamental way ([31]). But also from a more general methodological or philosophical point of view, fundamental criticism can be stated: Due to the open nature of the collection of heuristics propagated in most current accounts (i.e. whenever a phenomenon cannot be covered or described by an existing heuristics, a new one specifically fit to the task is introduced), the possibility of falsification and refutation of modeling assumptions and theories is not guaranteed, and a (reasonable) completion of the model can neither be checked for, nor feasibly assumed at any point.

3.3 A Survey: Rationality in AI

This section wants to shed some light on the concept of rationality (broadly construed), the role and importance of rationality for AI as a research endeavor, and its current standing and status within the field of AI. For the sake of the argument, I consider a very broad notion of what "rationality" is taken to be, as a limiting parameter for an action/behavior to be considered as rational only demanding for compliance with one or another variant of the quite abstract "principle of rationality" (i.e., requiring solutions to a task 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).

In the following, initially a short perspective on the relation between rationality and intelligence in general will be given before presenting some arguments why AI actually should care about notions and theories of rationality. Concludingly, I will provide a concise overview of earlier efforts and results concerning rationality in the context of artificial intelligence.

3.3.1 Rationality & (Natural) Intelligence

In many, often folk-psychologically motivated accounts of human cognition, the notion of intelligence is mostly seen as closely intertwined with a concept of rationality. Whilst in many cases a certain form of intelligence is taken as a precondition for rationality and rational behavior, rational behavior and rationality also may be considered one possible indicator for an agent's intelligence (i.e. making some form of intelligence a necessary condition for rationality to arise). Often, both phenomena are even understood as following some direct proportionality: The more intelligent a subject is, the more rational its behavior is expected to be (and vice versa).

But here, folk psychology clearly deviates from standard theoretical accounts and models for human rationality: Whilst the latter normally do not take individual factors of the subject (as, e.g., its level of intelligence, or overall capabilities) into account, the above stated conception – by making the level of rationality also dependent on the level of intelligence – clearly does.

An account which seems to be close in spirit to the assumption underlying this non-standard notion can also be found in scholarly study of rationality, namely in the field of decision theory. In [18], Gilboa (also see an earlier definition in [19]) defines rationality as a concept that crucially depends on the subject that is executing the reasoning and (allegedly) rational behavior: "*(...) a mode of behavior is rational for a given decision maker if, when confronted with the analysis of her behavior, the decision maker does not wish to change it.*". The consequences of this seemingly simple definition are tremendous: Rationality becomes subject-centered, as now what is considered as rational might vary with the population in question. If the decision maker does not understand the analysis, or why her behavior is not judged as rational, she cannot be judged irrational for not complying with the alleged norm of rationality. If limited cognitive capacities do not allow the reasoner to understand the rules he should follow in his reasoning, but would always make him

take the same decision again, he has to be called rational – although his behavior in that very moment might strike an observer as utterly irrational.

Skeptics now might be tempted to put forward the claim that, once one accepts such a notion of rationality, from that moment onwards there is no more difference between intelligence and rationality, but both notions collapse into one. To me, this objection seems mistaken: Although rational behavior might be witnessed as a (maybe even measurable) manifestation of intelligence, the effects of intelligence still do not have to be limited to rational behavior. It is commonplace that there are many different aspects to the overall notion of intelligence, an idea (amongst many others) going back to work by Howard Gardner. In [16], he lists at least eight abilities that might be considered facets of intelligence: spatial, linguistic, logical-mathematical, bodily-kinesthetic, musical, interpersonal, intrapersonal, and naturalistic. Clearly, some of the tasks and challenges addressing one or several of these abilities can also be subsumed by what a notion of rationality, built upon the aforementioned principle of rationality, can cover. Still, it should also become obvious that this coverage will not be complete, making intelligence the more general, by far more diverse and holistic conception.

3.3.2 Why Should AI Care About Rationality?

Artificial intelligence offers a wealth of concepts and notions called rational or claimed to reflect rationality. Treating rationality within a framework of artificial intelligence actually may have good reasons:

On the one hand, subscribing to a stance close to the human-like intelligence or "strong AI" research programs, modeling (human-style) rationality clearly has to be considered one of the milestones for reaching the overall goal of a "truly" intelligent AI system. Consider the example of what now is known as the "Artificial General Intelligence" (AGI) movement (cf. e.g. [37]): As explained there, an AI research project, in order to qualify for being an "AGI project", amongst others has to "*(...) be based on a theory about 'intelligence' as a whole (which may encompass intelligence as displayed by the human brain/mind, or may specifically refer to a class of non-human-like systems intended to display intelligence with a generality of scope at least roughly equalling that of the human brain/mind).*" Still humans and their performance and capabilities (although not without any alternative) stay the main standard of comparison for the targeted type of system – which in turn of course also assigns a crucial role to modeling and implementing human-style rational behavior. Similarly, a test for an artificial system's behavior in rationality tasks can serve as a crucial sub-task of a modernized decomposition [1] of Turing's famous test for machine intelligence [35].

On the other hand, staying closer to the ideas of specialized AI accounts, having a feasible concept of rationality within a system would allow agents to interact with humans in a more natural way, as well as to predict humans' expectations and interactive behavior. Applications for these capabilities would be manifold, ranging from decision support systems, through trading agents, to the use of intelligent agents in

flight control scenarios. Also, from a metaperspective, AI accounts of rationality could provide inspiration, modeling tools and testbeds for theories of rationality arising within related disciplines (cf. also Sect. 3.4).

3.3.3 Earlier Work & Status Quo

Several prominent researchers within the field of artificial intelligence explicitly addressed the relation between AI and rationality, as well as the role rationality can play within artificial intelligence. For instance, in [32], Russell elaborated on the relation between rationality and intelligence in an AI context. Starting out from the very premise of AI itself, the assumption that the understanding and (re)creation of intelligence is possible, Russell shows the importance of having a notion of intelligence that is precise enough to serve as a basis for both, the cumulative development of general results, as well as robust systems, before he considers rational agency as a candidate for fulfilling this role (also outlining the history and development of the corresponding concepts of rationality). And also Doyle [12] – claiming that a theory of rationality might at some point equal mathematical logic in its importance for mechanizing reasoning – provides a concise survey of work at the intersection between the economic theory of rationality and AI, offering insights into how basic notions of probability, utility and rational choice (together with some pragmatic considerations) can influence the design and analysis of reasoning and representation systems.

Moreover, it should be noted that certain conceptions of rationality have played an important role at different turning points within the development of artificial intelligence as a scientific field, sometimes emerging from new paradigms within AI, sometimes directly contributing to the creation of new stances and perspectives. Following the lines of a more detailed treatment in [9] (where rationality is regarded in a very general way as "*reason-governed behavior*"), for example the following research programs and key questions with direct impact from and/or on research in rationality can be named:

- Robotics: Can reasoned action be explained without making appeal to inner, form-based vehicles of meaning (i.e. are internal representations perhaps superfluous)? And in consequence, can there be something like representation-free rationality – can phenomena like what we consider "deliberative reasoning" or "abstract thought" be explained by a complex of reflex-like mechanisms alone?
- Global reasoning: How can "non-classical" forms of reasoning (e.g. abductive or analogical reasoning) which are clearly present in humans as real-world rational agents be accounted for in AI systems (where still deductive reasoning as the dominant paradigm)?
- Heuristics: Given the already aforementioned conceptions of fast and frugal heuristics [17] as real-world mechanisms of decision-making and reasoning, how must the principal ideas and understanding underlying most approaches to mechanizing reasoning (and thus also to mechanical rationality) be altered and adapted? Which techniques and AI paradigms are best fit to implement such heuristics-based mechanisms?

Concerning the status quo at the present moment, the different families of models for rationality listed in Sec. 3.2 (partly with exception of the heuristics-based approach) can actually be found within existing AI systems and theories. Still, when having a closer look at the latter, it shows that the underlying notions of rationality have stayed close to their fields of origin, having been adapted only to a minimal degree (if at all). Also, possible deficits and shortcomings have been brought along without questioning or fixing.

In consequence, AI systems e.g. mostly fall short in tasks such as predicting or exhibiting behavior resembling human-like rationality, a crucial need in all domains concerned with close interaction/cooperation between a human user and an AI system. Here, examples are numerous, ranging from rational agents communication for cooperative dialogues [34] to adaptive and cooperative wheelchairs [15].

3.4 The Goal: Rationality for AI

Recently, some researchers in cognitive science and decision theory are questioning the completeness and suitability of the classical approaches to rationality. For example, having a position not that different from Gilboa's view (sketched in Sect. 3.3.1), also Kokinov challenges traditional views on rationality in [23]: Initially observing that rationality fails as both, descriptive theory of human-decision making and normative theory for good decision-making, Kokinov concludes that the concept of rationality as a theory in its own right ought to be replaced by a multilevel theory based on mechanisms and processes involved in decision-making. He demands the classical concept of utility making to be rendered as an emergent property (as should the concept of rationality itself), emerging in most (but not all) cases.

One possible consequence of this stance is the proposition to use humans as gold-standard for actually existing rational agents, and consequently base models of rationality and rational behavior on cognitive capacities (as, for instance, analogy-making, cf. e.g. [2]). Clearly, this brings along a fundamental shift in the type of theory, aiming not for theories of normative nature, but trying to build a positive theory of human rationality (as e.g. also mentioned in a side remark in [5]): Until now, almost every theory of rationality put its emphasis on providing a framework for postdictively deciding whether an already taken action or decision ought to be considered rational or not. Contrastingly, a positive theory would focus on the predictive part of its model of rationality, aiming at feasibly predicting a rational agent's behavior and decisions when being provided with a precise-enough description of all information relevant for the respective situation (a stance which from my point of view also seems way closer to application scenarios and settings within AI).

First theoretical and empirical studies on using computational cognitive systems for these modeling tasks seem to provide support for the practical possibility and valuable applicability of approaches aiming at also taking cognitive aspects of humans into account when creating models of (human-style) rationality: In work on the JUDGEMAP system [27] for judgment and choice, particularities and characteristics which normally are taken as typical for humans could be reproduced in the

system's behavior, and considerations concerning applications of the computational analogy engine HDTP to rationality tasks (cf. e.g. [3]) offer new solution strategies for long-standing challenges for the standard models of rationality.

Following these examples, additionally being inspired from within AI e.g. by [33], I argue for establishing work on rationality as a proper, coordinated research program within artificial intelligence, not only limiting its focus to taking theories of rationality from respective neighboring disciplines and adapting them so that they can be applied within an AI framework, but to actively pursue research on (artificial) rationality on its own.

The advantages of such an approach seem significant: From an AI-internal perspective, it allows for approaching rationality within artificial intelligence from a more holistic point of view, possibly amalgamating existing accounts into theories and conceptions more suitable for the use in artificially intelligent systems. Here, early examples can, e.g., already rudimentarily be found in the development of probabilistic dynamic epistemic logics [24] (an attempt at integrating probabilistic and logical views of rationality), or in the conception of "algorithmic rationality" [22], which offers a framework for including computational costs in otherwise game-theoretical notions of rationality. The latter also may serve as an example for a second advantage of a properly AI-oriented notion of rationality: Most classical models and theories of rationality do not take into account computationally crucial factors as e.g. the complexity of a reasoning procedure. This might still be justifiable and legitimate in a field like economics or maybe even psychology, but shows to be a major problem when talking about rationality in an AI context. As AI this far necessarily and always is dealing with some implementation of a Turing machine, complexity issues do play a crucial role when implementing and emulating any kind of intelligence or cognitive capacity at any level, and the tractability of the used methods and algorithms becomes a key issue (for a related perspective from the field of Cognitive Science, cf. also [30]). When establishing work on rationality as a proper program within AI, one intuitive starting point would therefore be to revisit the traditional accounts and models of rationality developed over the decades, and perform a rigorous complexity and tractability analysis on them. Doing this in a coordinated, collaborative way within an overall research program can reasonably be expected to be beneficial in several ways, amongst others allowing for the re-use of results and insights between different sub-endeavors, the immediate uptake of new ideas and conceptions into new models, frameworks and possibly even paradigms by neighboring projects and groups within the overall rationality AI program, but also creating the possibility to directly test and evaluate newly constructed models and theories against the then existing benchmarking infrastructure (on both, a conceptual and a computational level).

Finally, taking all these aspects into account, the proposed approach would thus also allow for providing feedback to neighboring disciplines, directly guiding further research in rationality concepts for AI, and could thereby play an integrative role similar to the artificial general intelligence program within the development of "strong AI" (cf. e.g. [37]).

From a more high-level point of view, it gives a basis for closer cooperation with neighboring fields, allowing for more and deeper (also not only unidirectional, but bidirectional) interaction. Such joint efforts might span a very wide range of scenarios:

- Concrete project-based joint ventures for example in human-computer interaction: In [7], Butterworth and Blandford state that an approach to reasoning about interactive behavior can be based on the assumption that computer users are rational, and that the behavior of the interactive system as a whole results from the rational behavior of the users in combination with the programmed behavior of the system's parts. Of course, the reliability of such an approach crucially depends on the quality of the rationality model, and the latter's suitability for such a use – where specifically designed models of rationality, deviating from the classical paradigms and (up to a high degree) reliably predicting human rational behavior, can be expected to have major benefits and advantages over the classical postdictive theories and accounts.
- Rather specific applications of AI methods and systems for mitigating shortcomings of theoretical frameworks in economics: In [11], whilst still maintaining a rather orthodox view on economic rationality theory, Dixon also elaborates on possible contributions of the field of artificial intelligence to the study of rationality. He sketches two examples where AI might play a significant role, one of them being disequilibrium situations (i.e. situations in which agents in their behavior deviate from the classical notion of equilibrium and thus seemingly err in their actions), and the second one being strategic environments (i.e. situations and contexts in which agents can increase their obtained utility by behaving seemingly non-optimal). Whilst economic theory faces great difficulties in modeling both phenomena, Dixon is rather confident that AI-inspired techniques could offer new insights and results.
- Inspiration, modeling tools and testbeds for theories of rationality within other disciplines: In [14], Frantz traces the origins of Herbert Simon's ideas about limited or bounded rationality back to Simon's work in the field of AI, and provides evidence that the former only could come about on the basis of Simon's previous experiences with the latter.
- Direct new contributions to longstanding concepts of rationality, as for instance the already mentioned notions of algorithmic rationality [22] or probabilistic dynamic epistemic logic [24].
- Fundamental research on human-like rationality conducted together with researchers on the cognitive science-side of system design and modeling, epistemology and philosophy of cognition: Here, Pollock's OSCAR Project (cf. e.g. [28]) can serve as an example. The project's aims were twofold, on the one hand, a general theory of rational cognition should be constructed, on the other hand, also an artificial rational agent implementing that theory should be built, resulting in a joint project in (at least) philosophy of cognition and AI. The basic idea of this particular approach was to offer the philosophical side a possibility of testing the correctness of the corresponding theory of rational cognition by applying it to concrete examples via the agent, whilst AI should be provided with a

general characterization of what it is to make rational decisions and draw rational conclusions (i.e. a general theory of rational cognition) from the philosophical side.

The success (or lack thereof) of such a rationality program within AI could at least be measured against two dimensions, namely the overall impact within AI and the impact within the classical "rationality disciplines" (i.e. psychology, economics, cognitive science). If the outcomes of associated research endeavors would contribute to coming closer to a proper and more natural (in the sense of flexible, adaptable or behaviorally adequate) model of rational behavior within artificial agent populations, or should allow for artificial systems predicting human decision-making and judgment in a more reliable and feasible way than actual architectures do, certain success of the research program could not be neglected. And also concerning the impact and consequences for topically related disciplines, a similar argument can be made: If insights resulting from the work done in AI (as, e.g., concerning the already mentioned complexity studies of existing models of rationality) can help clarifying open questions concerning the suitability and applicability of existing frameworks and theories, or can contribute to constructing better and maybe even more cognitively adequate models, the efforts spent would have to be judged worthwhile. And, although possibly more complicated to measure, also a third dimension of evaluation might be considered, namely the bridging of gaps and the creation of new connections between AI and some of its conceptually related neighbors in the world of academic disciplines, for example re-establishing some more of the original (programatic, conceptual and even methodological) links between AI, cognitive science and (cognitive) psychology.

3.5 An Outlook: Rationality Through AI

In this penultimate section, I very briefly want to sketch a few essentials and guidelines of what I consider a concept and theory of (human-style) rationality to which AI could contribute greatly (a more detailed elaboration can for example be found in [4]). This notion fundamentally deviates from most "standard" accounts of rationality, as it aims at a positive rather than normative type of model in the first place (deriving the normative dimension from the positive one once the latter has been constructed) and does not make any a priori commitments to particular formalisms or modeling techniques.

From my point of view, rationality in a human context clearly has to be considered as a subject-centered notion, also taking into account the particular capabilities and limitations of the respective agent. This position also includes the claim that there is no use establishing models and norms which can never be implemented or fulfilled by human reasoners due to limitations of the agent or of its environment (again, cf. e.g. [30]). Also, the main aspect of theories and models of rationality should be their use and applicability as positive theories, and not as mere normative or postdictive explanatory accounts. Such a model is intended to provide a reliable prediction of human rational behavior and decision-making when being supplied

with the information the human reasoner can access at the moment of reasoning (and with that information only).[2] Thirdly, there does not have to be any kind of commitment to a particular modeling paradigm or technique, but an integrative approach including and integrating different formalisms and approaches to modeling is perfectly admissible.

Provided with these guiding principles, it should become clear why AI can and should play an important role in the development of the respective theory of rationality. All three major demands, i.e. the awareness of the limits and boundaries of an agent, the request for applicability as predictive model, and the acceptance of systems which integrate different formal approaches and paradigms, are already popular (if not even commonplace) in important parts of AI as a field. Therefore, this expertise and experience could be very beneficial in the development of such a positive, subject-centered notion of rationality, whilst on the other hand AI itself could greatly profit from being provided with results and theoretical outcomes from this endeavor.

I want to conclude this section with a short statement about what rationality in AI from my point of view does not have to be or provide: Theories of rationality, although hopefully being of the type just sketched in the previous paragraph, do not have to re-construct or re-create human rationality (or even human intelligence, bringing along rationality as a by-product) on a level similar to the "algorithmic level" in Marr's Tri-Level Hypothesis of information processing [10], but can stay limited to the "computational level". Although inspiration might be taken from how humans actually perform rational decision-making and judgment, insights from respective studies should not have mandatory character, but are at most to be taken as recommendations (that can, and most likely will, be ignored in quite some cases). This conception also brings along a perspective different from the basic approach to AI that seems to underlie the thoughts presented in Chap. 4 of this book. Whilst some researchers take it as a necessity to use the proper "human model" and the proper "human implementation", thus trying to re-implement particularly human faculties and capacities when trying to build an AI (seemingly aiming for a complete match at least on the computational level and on the algorithmic level in Marr's hierarchy), I take a stance much closer to the intuitions also shining through in the Turing Test [35]. For me, the important part is to get the behavior right, and not so much the way this comes about – or, quoting a philosopher friend, phrasing it in an overly simplified but quite appealing way: "*Rationality is, what rationality does.*"

3.6 Conclusion

We have now seen a short assessment of the history and the actual status of the concept of rationality within AI, have considered the possible merits, value and

[2] My approach at this point has close connections to many of the ideas and insights underlying the notion of "ecological rationality" ([29]) discussed in economics and psychology. Also, both accounts share the general fundamental criticism concerning the classical normative approach to rationality.

consequences of work on models and theories of rationality using AI techniques and methods, and have also had a quick look at a non-standard (positive instead of normative) notion of rationality that seems more useful to the purposes of AI than most classical accounts do.

Summarizing, I once again want to put forward the claim and demand that targeted research on theoretical and applied aspects of rationality within AI should be conducted, using methods from AI and, in the long run, aiming at solving some of the fundamental problems the overall artificial intelligence program tries to answer. Whilst research on rationality has – for historical, but also for social reasons within scientific communities – by now mostly been conducted almost exclusively in the fields of economics, decision theory and psychology, I am convinced that a proper research program within AI could be fruitful not only for the purposes of AI itself (here also allowing for new approaches and takes on long-standing challenges), but also would provide valuable feedback and new perspectives for the already established "players in the field of rationality".

Acknowledgements. The author wishes to thank Maricarmen Martínez Baldares for her helpful recommendations and comments concerning some of the mentioned topics. Also, thanks go to Frank Jäkel and the members of the AI Research Group at the Institute of Cognitive Science for quite some valuable discussion and exchange of ideas. Finally, I owe a debt of gratitude to Wendy Wilutzky for coining the concluding quote of Sect. 3.5.

References

1. Besold, T.R.: Turing Revisited: A Cognitively-Inspired Decomposition. In: Müller, V.C. (ed.) Theory and Philosophy of Artificial Intelligence. SAPERE (to appear, 2012)
2. Besold, T.R., Gust, H., Krumnack, U., Abdel-Fattah, A., Schmidt, M., Kühnberger, K.: An Argument for an Analogical Perspective on Rationality & Decision-Making. In: van Eijck, J., Verbrugge, R. (eds.) Proceedings of the Workshop on Reasoning About Other Minds: Logical and Cognitive Perspectives (RAOM 2011), CEUR Workshop Proceedings, Groningen, The Netherlands, vol. 751, CEUR-WS.org (2011)
3. Besold, T.R., Gust, H., Krumnack, U., Schmidt, M., Abdel-Fattah, A., Kühnberger, K.U.: Rationality Through Analogy - Towards a Positive Theory and Implementation of Human-Style Rationality. In: Troch, I., Breitenecker, F. (eds.) Proc. of MATHMOD 12, Vienna (2012)
4. Besold, T.R., Kühnberger, K.U.: E Pluribus Multa In Unum: The Rationality Multiverse. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society, Cognitive Science Society, Austin (2012)
5. Binmore, K.: Rational Decisions. Princeton University Press (2011)
6. Broadie, S., Rowe, C. (eds.): Aristotle Nicomachean Ethics: Translation, Introduction, and Commentary. Oxford University Press (2002)
7. Butterworth, R., Blandford, A.: The principle of rationality and models of highly interactive systems. In: Sasse, M.A., Johnson, C. (eds.) Human-Computer Interaction – INTERACT 1999, IOS Press (1999)
8. Byrne, R.: Suppressing valid inferences with conditionals. Cognition 31(1), 61–83 (1989)

9. Clark, A.: Artificial Intelligence and the Many Faces of Reason. In: The Blackwell Guide to Philosophy of Mind. Blackwell (2003)
10. Dawson, M.: Understanding Cognitive Science. Blackwell Publishing (1998)
11. Dixon, H.: Some Thoughts on Economic Theory and Artificial Intelligence. In: Surfing Economics: Essays for the Enquiring Economist, Palgrave (2001)
12. Doyle, J.: Rationality and its roles in reasoning. Computational Intelligence 8(2), 376–409 (1992)
13. Evans, J.: Logic and human reasoning: An assessment of the deduction paradigm. Psychological Bulletin 128, 978–996 (2002)
14. Frantz, R., Simon, H.: Artificial intelligence as a framework for understanding intuition. Journal of Economic Psychology 24, 265–277 (2003)
15. Galluppi, F., Urdiales, C., Sandoval, F., Olivetti, M.: A study on a shared control navigation system: human/robot collaboration for assisting people in mobility. Cognitive Processing 10(2), 215–218 (2009)
16. Gardner, H.: Frames of Mind: The Theory of Multiple Intelligences. Basic Books, New York (1983)
17. Gigerenzer, G., Hertwig, R., Pachur, T. (eds.): Heuristics: The Foundation of Adaptive Behavior. Oxford University Press (2011)
18. Gilboa, I.: Questions in decision theory. Annual Reviews in Economics 2, 1–19 (2010)
19. Gilboa, I., Schmeidler, D.: A Theory of Case-Based Decisions. Cambridge University Press (2001)
20. Griffiths, T., Kemp, C., Tenenbaum, J.: Bayesian Models of Cognition. In: The Cambridge Handbook of Computational Cognitive Modeling. Cambridge University Press (2008)
21. Halpern, J.Y.: Beyond Nash Equilibrium: Solution Concepts for the 21st Century. In: Proc. of the 27th Annual ACM Symposium on Principles of Distributed Computing (2008)
22. Halpern, J.Y., Pass, R.: Algorithmic rationality: Adding cost of computation to game theory. ACM SIGecom Exchanges 10(2), 9–15 (2011)
23. Kokinov, B.: Analogy in decision-making, social interaction, and emergent rationality. Behavioral and Braind Sciences 26(2), 167–169 (2003)
24. Kooi, B.P.: Probabilistic dynamic epistemic logic. Journal of Logic, Language and Information 12, 381–408 (2003)
25. Mill, J.S.: On the definition of political economy, and on the method of investigation proper to it. In: Essays on Some Unsettled Questions of Political Economy, 2nd edn. Longmans, Green, Reader & Dyer (1874)
26. Osborne, M., Rubinstein, A.: A Course in Game Theory. MIT Press (1994)
27. Petkov, G., Kokinov, B.: JUDGEMAP - integration of analogy-making, judgement, and choice. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci), pp. 1950–1955 (2006)
28. Pollock, J.: Twenty Epistemological Self-profiles: John Pollock (Epistemology, Rationality and Cognition). In: A Companion to Epistemology, pp. 178–185. John Wiley and Sons (2010)
29. Rieskamp, J., Reimer, T.: Ecological rationality. In: Encyclopedia of Social Psychology, pp. 273–275. Sage, Thousand Oaks (2007)
30. van Rooij, I.: The tractable cognition thesis. Cognitive Science 32, 939–984 (2008)
31. van Rooij, I., Wright, C., Wareham, H.T.: Intractability and the use of heuristics in psychological explanations. Synthese (2010) (published online: November 11, 2010)
32. Russell, S.: Rationality and intelligence. Artificial Intelligence 94, 57–77 (1995)

33. Russell, S.J., Wefald, E.: Do the right thing - studies in limited rationality. MIT Press (1991)
34. Sadek, M.D., Bretier, P., Panaget, F.: ARTIMIS: Natural dialogue meets rational agency. In: Proceedings of IJCAI 1997, pp. 1030–1035. Morgan Kaufmann (1997)
35. Turing, A.: Computing Machinery and Intelligence. Mind LIX (236), 433–460 (1950)
36. Tversky, A., Kahneman, D.: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. Psychological Review 90(4), 293–315 (1983)
37. Wang, P., Goertzel, B.: Introduction: Aspects of Artificial General Intelligence. In: Goertzel, B., Wang, P. (eds.) Advances in Artificial General Intelligence - Proc. of the AGI Workshop 2006, Frontiers in Artificial Intelligence and Applications., vol. 157, pp. 1–16. IOS Press (2007)
38. Wason, P.C., Shapiro, D.: Natural and contrived experience in a reasoning problem. The Quarterly Journal of Experimental Psychology 23(1), 63–71 (1971)