Report on Rational Analysis

by Felix Schmitt

Darmstadt

Human Sciences
Department
Institute for Psychology

Contents

1 Introduction

When it comes to Cognitive Science, nowadays most researchers dedicate themselves to one of three approaches of cognitive modeling: Parallel Distributed Processing, also known as Connectionism, Cognitive Architectures or Rational Analysis. While the first tries to understand the human mind by regarding it as a highly complex parallel processing unit and comparing it with neural networks, proponents of Cognitive Architectures mainly focus on how different components of a complex system interact to lead to intelligent behavior. While both of these approaches appear similar at first glance, there has been a long history of debates on which is the better way to solve the mystery of human intelligence. Although either of them have a considerable amount of success stories, many proponents still believe the opposing perspective to be wrong and misleading.

In contrast to this, the third approach, Rational Analysis, comes from an entirely different direction: Rather than approximating the mind via a given framework, it tries to find a model that ideally solves a specific task and then modifies the model to match empirical data from human experiments. Although this is most prominently done with Bayesian methods, the approach does not restrict itself to one family of algorithms, but is rather able to solve the same task via multiple different techniques, of which many can be possible candidates of what is really happening within the brain. This can be seen as a detriment and an advantage at the same time: On one hand, Rational Analysis lacks the ability to clearly determine a concrete model on how cognition *actually* works, on the other it is able to find a wide range of likely explanations outside the scope of Connectionism and Cognitive Architectures. Although researchers from both these fields have moved to combining their approaches [21], proponents of Rational Analysis regularly make use of them, often also mixing them with techniques from entirely different fields of research [19]. In the following, I will describe this process in detail, using a reverse-engineering approach that works it's way from the higher cognitive processes to the underlying mechanisms in the brain.

2 Summary of the approach

As the name already tells, Rational Analysis assumes that an agent performs a given task in a somewhat rational way, optimizing it's performance by choosing an action that leads to a maximal reward using minimal effort. This assumption is supported by evolution, as it fits the idea of an organism slowly converging to optimal adaption to it's environment [3]. What is particularly important on this is the role of the environment. Unlike in experimental setups, the natural environment of organisms is almost always changing, resulting in uncertainty for the agent within it. To still be able to predict future states of the environment, an agent has to use probabilities, so it can act accordingly to the most likely future state, while also being prepared for other less likely states.

With the Bayes-Theorem leading to optimal results for most probability computations, it (or some variation of it) is therefore likely to be used by the brain. This, however, does not mean that all of human cognition follows Bayesian rules, neither that humans can understand all their tasks and mental representations in a Bayesian way. As Tenenbaum pointed out: "The claim that human minds learn and reason according to Bayesian principles is not a claim that the mind can implement any Bayesian inference. Only those inductive computations that the mind is designed to perform well, where biology has had time and cause to engineer effective and efficient mechanisms, are likely to be understood in Bayesian terms" [26]. As examples where the Bayesian approach led to promising results, he names perception, language, memory and sensorimotor systems [28, 8, 23, 25, 17]. On the other hand, challenges relatively new to mankind, like computations involving multiple probabilities, are subject to severe biases and can hardly be solved by most people [27]. Fascinatingly, replacing these probabilities with natural numbers (e.g. 5 out of 1000) leads to a drastic increase in human performance, which can be attributed to the brain having learned this in prehistoric times [7].

Being aware of the uncertainty of the environment and the resulting necessity of probabilistic methods, as well as the assumption that agents try to behave optimal, we can now formulate an ideal observer for a given task. The ideal observer serves as a starting point from which we can deduce the way a human would solve this task, while also being an excellent comparison for actual human performance. Obviously humans are rarely capable of optimally solving a somewhat complex task, which not only stems from the noisiness of the environment, but also that of the humans perception and it's processing. Therefore, Rational Analysts introduce a wide arrange of tweaks to the ideal observer, to more closely match it's performance with human data. In their 2016 paper [29], Zednik and Jäkel list four different forms of tweaks, that Rational Analysts regularly use to achieve the wanted behavior:

1) The added-limitations tweak

Due to the innate noise that comes with biological neurons, humans are unable to accurately perceive the exact state of their environment, resulting in imperfect representations in their brains. Therefore it is common to also limit the perception and simulation capabilities of the ideal observer model, to more closely match the performance of humans.

2) The different-environment tweak

Due to the human data always being recorded in somewhat artificial experiments, the properties of the environment rarely match those of the subjects natural surroundings. This shows itself in apparent biases of the participants, which can often be deduced to their performance being optimized to their natural environment. To account for this, the model can be tweaked to match the performance of a specific participant, possibly even revealing which properties of their natural environment they have adapted to.

3) The subjective-optimality tweak

Unlike an ideal observer, human participants always have subjective needs and preferences, distorting their maximal potential performance. These can range from the preference of an option when unsure, to fatigue or even the urge to blink. Being particularly hard to detect or avoid, these biases can mostly just be assumed and then be added to the model for a specific participant.

4) The suboptimality tweak

Lastly, if none of the previous tweaks were able to fit the data sufficiently, it has to be assumed that the performance of the participant was simply suboptimal. As already stated before, humans are far from able to solve every task they are confronted with in an optimal way, whether it is due to insufficient training, unoptimized performance already achieving their goal, or simply too high difficulty of the task. In this case, the ideal observer model has to be impaired, for example by making the used values more vague or even introducing artificial noise into the computation.

After applying a combination of these tweaks, the researcher has fit the data of the ideal observer model to that of the human participants. This already leads to some interesting conclusions: Depending on the types of applied tweaks, one can already assume which factors effect human performance on the given task, while also learning how the experiment could be modified to further optimize it. It is very important to be aware, that the applied tweaks do not necessarily correspond to the actual biases underlying the human data, but rather are just assumptions what the researcher believes to have led to the discrepancies between human and ideal data. In most cases though, there is sufficient evidence for the chosen tweaks to at least be a very likely cause of the bias.

Until this point, the Rational Analysis-approach has only touched the computational level of Marr's three levels of analysis. Several researchers stop here, beginning to draw conclusions from their proposed models and publishing their findings. As criticized by many, this leads to more or less unfalsifiable and purely hypothetical theories of cognition [16].

Therefore, Zednik and Jäkel suggest to move towards Bayesian reverse-engineering. This approach uses the conclusions from the Rational Analysis as a mere starting point for the following research. Being a top-down model, it takes the explanations for the higher processes of cognition from the Rational Ananlysis, which then give rise to the underlying mechanisms that produce the input for the computational level.

For this, the researcher has to find an algorithm that fits the previously tweaked ideal observer. As one can imagine, there is a sheer endless amount of possible algorithms the researcher has to choose from. Here, Zednik and Jäkel refer to Simons [24] idea of using heuristic strategies, to reduce the size of the hypothesis-space. Similar to the tweaks, they propose a list of five different heuristics one should follow when trying to find a suitable representation for the algorithmic and the implementational level:

1) The push-down heuristic

What has previously been implicitly stated on the computational level, will now be explicitly formulated on the algorithmic level. More precisely, this means the researcher now tries to find an algorithm, that fits the hypothetical procedure specified by the tweaked ideal observer model. As the name of the heuristic already suggests, this algorithm is practically pre-determined by the chosen computational model, therefore "pushed down" from it. In practice, this means that for the same exact task, different theories of the computational level lead to different representations on the algorithmic level. It is important to note, that different computational theories can be formulated for the same ideal observer model. For example, one can either assume that a discrimination task is done by applying Bayes rule to determine an objects category, or by comparing each object to a simple criterion which then classifies it's category. Both approaches can result in identical output, but lead to completely different models for the computational level and therefore also the algorithmic level.

2) The tools-to-theories heuristic

In contrast to the push-down heuristic, the tools-to-theories heuristic does not directly infer the algorithm from the computational level, but rather takes mechanisms from different fields of research as inspiration for the algorithmic representation. In most cases, these stem from machine learning, artificial intelligence or statistics and were initially applied in areas like medicine, meteorology or image analysis [2]. These “tools” are then modified to fit the computations of the ideal observer model. Due to the majority of AI and machine learning algorithms solving problems previously been worked on by humans, they naturally serve as good candidates for problem solving in the human mind. Much like for the push-down heuristic, the algorithms chosen with this heuristic are from a huge scope of possible solutions and are greatly influenced by given factors: What has been the chosen computational model for the push-down heuristic, is often the research domain of the executing scientist for the tools-to-theories heuristic. If the researcher mainly works with machine learning, he is more likely to pick a machine learning algorithm, than for example a statistician.

3) The unification heuristic

Unlike the previous heuristics, who tried to find an algorithm from the infinite scope of possible solutions, this heuristic helps to pick out one of a few likely candidates. It does so by preferring the algorithm with the widest range of other applications for human cognition. As Colombo and Hartmann [9] have argued, this does not necessarily relate to a unifying approach being more mechanically feasible, but can also be due to its mathematical properties fitting a multitude of cognitive challenges.

4) The plausible-algorithms heuristic

This heuristic takes previous findings in psychology and cognitive science into account, by excluding algorithms contradicting these findings. The easiest example would be algorithms that are so computationally demanding, they would not generate a meaningful output in the time the brain actually does. With this heuristic, it is also possible to apply algorithms that have proven useful in different psychological research areas to the cognitive process one is examining.

5) The possible-implementations heuristic

While the previous heuristics all serve to get from the computational level to the algorithmic level, the possible-implementations heuristic reaches even further, into the implementational level. Much like the plausible-algorithms heuristic, it evaluates possible algorithms based on previous findings. In this case, these findings stem from neuroscience and brain imaging. With these particular areas of research still being in their infancy, they are only able to provide approximate data, which therefore can not be used to draw direct conclusions. As the other heuristics though, it can be very useful to reduce the pool of possible candidates and even infer very elaborated theories [12, 4].

Now, we have a step by step procedure from the top to the bottom of Marr's three levels: Starting on the computational level, the researchers formulate an ideal observer model for their investigated task and tweak it until it matches human data. From there on, they apply a variety of heuristics to pin down likely representations on the algorithmic and implementational level. As stated before, Rational Analysis is unable to confirm that the model chosen at the end is actually corresponding to the cognitive processes involved in the task. It is by nature rather speculative and uses many assumptions on all of Marr's levels. Due to these assumptions stemming from a variety of highly precise evaluation steps, the suggested models are mostly still excellent proposals, if not guiding the way of future research. Especially with brain imaging advancing [1, 22], the possibility of confirming these theories on the implementational level is coming closer and closer.

Worth emphasizing, the procedure of Rational Analysis was explained without focusing on particular algorithms, to show it's general applicability as a research method. In practice, most proponents of it mainly use Bayesian approaches, because they:

1) Almost impose themselves as ideal observers in an uncertain environment, resulting in optimal performance for probability theory

2) Can easily be adapted to fit the tweaks of the ideal observer model

3) For many tasks are induced by almost all the heuristics, especially the unification and the tools-to-theories heuristic [11, 15]

3 Critiques and Answers

With their 2011 paper, Jones and Love [16] formulated one of the most common points of critique on Rational Analysis: “Much of this research aims to demonstrate that cognitive behavior can be explained from rational principles alone, without recourse to psychological or neurological processes and representations.” In other words, Rational Analysis only draws hypothesis on the computational level, ignoring their algorithmic- and implemetational-level foundations. According to them, this results in unfalsifiability, vagueness and normativeness, lacking evidence on actual observations. Or, as Chandrasekaran [6] framed it on his comment on Andersons 1991 paper [3], Rational Analysis does not explain “how to *create* mind-like entities”. While this holds true for a variety of publications, especially older ones like Andersons, more recent Rational Analysis-models are often based on neurological and empirical findings [4, 12], explaining how these algorithms could be implemented in the brain. In particular by following the outline described by Zednik and Jäkel, researchers adapt their models to state-of-the-art psychological and neuroscientific data (see plausible-algorithms heuristic and possible-implementations heuristic). Comparing this to the approaches of Connectionism and Cognitive Architecures, one should combine the evidence for all of them on all three of Marr’s levels. While Connectionism is particularly strong on the implementational level, it rarely produces tangible explanations on the computational level. Cognitive Architectures on the other hand, struggle to find an implementational level representation, with the proposed symbolic mechanisms still being a topic of debate. Regarding this, it seems like Rational Analysis is not lagging behind, but is rather on par with it’s competing approaches in relation to Marr’s levels.

Another prominent point of critique is the apparent assumption of Rational Analysts, that all forms of cognition are computed in an “optimal” or “rational” way. In their response to Goodman et al. [13], Marcus and Davis argue that “strong, unwarranted claims for the optimality of performance are often made in the literature on Bayesian models” [20]. While this obviously halts true for some publications [10, 18], the majority of researchers uses the notion of optimality with caution. For example, both Tenenbaum and Anderson

[26, 3] emphasize the role of long-term biological adaptation, necessary for optimal behavior to even arise. It should also be noted, that the previously explained suboptimality tweak already contradicts Marcus' and Davis' point. Zednik and Jäkel even go as far as calling the "issue of rationality [...] a red herring". They reason that optimality and therefore ideal observers merely serve as a reference point "to navigate the space of computational-level hypotheses and to guide subsequent investigations at the algorithmic and implementational levels in a way that might even reveal suboptimal processes and mechanisms" [29]. With this in mind, it makes sense that ideal observers rarely match human data, rendering it unimportant if this is caused by humans naturally behaving suboptimal, or their behavior being optimized to latent properties.

Due to their flexible nature on the computational level, Rational Analysis models are often criticized as being unfalsifiable and arbitrarily chosen by the conducting researcher [5]. This point of critique is beautifully repelled by Griffiths et al. [14], who emphasize the distinction between a model and a theoretical framework. While a variety of models induced through Rational Analysis are in fact really unfalsifiable [3], this does not account for the overarching theoretical framework, Rational Analysis itself. They argue that such a framework serves as "a general perspective and a set of tools for making models" and is therefore by definition not falsifiable. In fact, the same critique could be applied to Connectionism and Cognitive Architectures, who are both also very flexible due to their high amount of degrees of freedom. For Connectionism they lay in the choice of architecture, learning algorithm, initialization, and training set, for Cognitive Architectures in the available production rules and the mechanisms selecting them. Neither of them are regularly criticized as being unfalsifiable or too flexible. In summary, it is always possible to find a few bad applications of a framework, but rarely correct to infer from them to the framework as a whole. Besides that, this criticism mainly focuses on the computational level. When for example following the Bayesian reverse-engineering process, as described by Zednik and Jäkel, researchers also formulate hypothesis for the algorithmic and implementational level. Much like their counterparts in Connectionism and Cognitive Architectures, these can be falsified with the typical psychological or neuroscientific procedures.

4 Personal Opinion

After reading a good amount of critiques on Rational Analysis, including heated back-and-forth discussions between proponents and critics of the approach, I get the impression that scientific resources are literally wasted here. I am aware of the fact that theories, especially extensive frameworks like Rational Analysis, should be thoroughly reviewed and verified, and that this process is necessary to solidify it's role as a useful scientific approach. But nonetheless, the repeated exchange of attacks and defenses of the approach seems more emotionally driven than like a scientific discourse to me. Both Jones and Love, as well as Marcus and Davis, delivered a fair amount of justified criticism, which then was in large parts rebutted by Rational Analysts like Goodman and Griffiths. At this point I do not see the necessity to continue criticizing the approach, especially by reframing previously used arguments and picking single examples of it to transfer them onto it as a whole.

It is not that I completely disagree with everything stated by the critics, in fact many points they make reflect my own worries about Rational Analysis. But as for pretty much everything, the truth lies somewhere in between the different paths. Disregarding Rational Analysis as useless to me is equally ignorant, if not even more, than declaring advances in Connectionism or Cognitive Architectures as valueless, because they do not fit in one's own preferred approach to cognitive modeling. Especially with Rational Analysis often merging them together, I think it is particularly promising to generate new findings in Cognitive Science. As almost all prominent proponents of Rational Analysis emphasize, they are not "advocating [it] to the exclusion of other approaches to cognition" [3]. It rather serves as a promising guideline, helping researchers from all schools of thought to determine likely implementations of the processes they study.

Considering the highly complex organization and interconnection of neurons in the brain, as well as the functional differences between the neurons themselves, it is safe to say that the current architecture of neural nets does not resemble what happens in the brain. However, this does not render Connectionism useless, as many experiments have also shown it’s power to explain a variety of cognitive processes. The same can be said for Cognitive Architectures and Rational Analysis. From this it is an obvious conclusion, that some combination of them, most likely including other approaches not yet discovered, are needed to explain the true fundamentals and mechanisms of the brain. With all three approaches still being far from optimized, I hope that in the future they evolve into a more intertwined strategy of solving the riddle of intelligence, that has accompanied us from our earliest days.

To close of, I want to cite a part of the response Griffiths et al. [14] gave to the critique of Bowers and Davis in 2012 [5]: “Different theoretical frameworks, such as Bayesian modeling, connectionism, and production systems, have different insights to offer about human cognition [...]. A connectionist model and a Bayesian model of the same phenomenon can both provide valuable information [. . .] and both could well be valid. The ultimate test of these different theoretical frameworks will be not whether they are true or false, but whether they are useful in leading us to new ideas about the mind and brain, and we believe that the Bayesian approach has already proven fruitful in this regard.”

Bibliography

[1] Juan Alvarez-Linera. “3 T MRI: Advances in brain imaging”. In: *European journal of radiology* 67.3 (2008), pp. 415–426.

[2] Samaneh Aminikhanghahi and Diane J Cook. “A survey of methods for time series change point detection”. In: *Knowledge and information systems* 51.2 (2017), pp. 339–367.

[3] John R Anderson. *Is human cognition adaptive?* na, 1991.

[4] Pietro Berkes et al. “Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment”. In: *Science* 331.6013 (2011), pp. 83–87.

[5] Jeffrey S Bowers and Colin J Davis. “Bayesian just-so stories in psychology and neuroscience.” In: *Psychological bulletin* 138.3 (2012), p. 389.

[6] B Chandrasekaran. “Mechanistic and rationalistic explanations are complementary”. In: *Behavioral and Brain Sciences* 14.3 (1991), p. 489.

[7] Valerie M Chase, Ralph Hertwig, and Gerd Gigerenzer. “Visions of rationality”. In: *Trends in cognitive sciences* 2.6 (1998), pp. 206–214.

[8] Nick Chater and Christopher D Manning. “Probabilistic models of language processing and acquisition”. In: *Trends in cognitive sciences* 10.7 (2006), pp. 335–344.

[9] Matteo Colombo and Stephan Hartmann. “Bayesian cognitive science, unification, and explanation”. In: *The British Journal for the Philosophy of Science* 68.2 (2017), pp. 451–484.

[10] Marc O Ernst and Martin S Banks. “Humans integrate visual and haptic information in a statistically optimal fashion”. In: *Nature* 415.6870 (2002), pp. 429–433.

[11] Sebastian Farquhar and Yarin Gal. “A unifying bayesian view of continual learning”. In: *arXiv preprint arXiv:1902.06494* (2019).

[12] József Fiser et al. “Statistically optimal perception and learning: from behavior to neural representations”. In: *Trends in cognitive sciences* 14.3 (2010), pp. 119–130.

[13] Noah D Goodman et al. “Relevant and robust: A response to Marcus and Davis (2013)”. In: *Psychological science* 26.4 (2015), pp. 539–541.

[14] Thomas L Griffiths et al. “How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012).” In: (2012).

[15] Nachikethas A Jagadeesan and Bhaskar Krishnamachari. “A unifying Bayesian optimization framework for radio frequency localization”. In: *IEEE Transactions on Cognitive Communications and Networking* 4.1 (2017), pp. 135–145.

[16] Matt Jones and Bradley C Love. “Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition”. In: *Behavioral and brain sciences* 34.4 (2011), p. 169.

[17] Konrad P Körding and Daniel M Wolpert. “Bayesian decision theory in sensorimotor control”. In: *Trends in cognitive sciences* 10.7 (2006), pp. 319–326.

[18] Konrad P Körding and Daniel M Wolpert. “Bayesian integration in sensorimotor learning”. In: *Nature* 427.6971 (2004), pp. 244–247.

[19] Vikash K Mansinghka et al. “Approximate bayesian image interpretation using generative probabilistic graphics programs”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 1520–1528.

[20] Gary F Marcus and Ernest Davis. “Still searching for principles: A response to Goodman et al.(2015)”. In: *Psychological Science* 26.4 (2015), pp. 542–544.

[21] Alexey Potapov et al. “Semantic image retrieval by uniting deep neural networks and cognitive architectures”. In: *International Conference on Artificial General Intelligence*. Springer. 2018, pp. 196–206.

[22] Pranoy Sahu and Nirmal Mazumder. “Advances in adaptive optics–based two-photon fluorescence microscopy for brain imaging”. In: *Lasers in medical science* (2020), pp. 1–12.

[23] Richard M Shiffrin and Mark Steyvers. “A model for recognition memory: REM—retrieving effectively from memory”. In: *Psychonomic bulletin & review* 4.2 (1997), pp. 145–166.

[24] Herbert A Simon, Patrick W Langley, and Gary L Bradshaw. “Scientific discovery as problem solving”. In: *Synthese* (1981), pp. 1–27.

[25] Mark Steyvers, Thomas L Griffiths, and Simon Dennis. “Probabilistic inference in human semantic memory”. In: *Trends in cognitive sciences* 10.7 (2006), pp. 327–334.

[26] Joshua B Tenenbaum et al. “How to grow a mind: Statistics, structure, and abstraction”. In: *science* 331.6022 (2011), pp. 1279–1285.

[27] Amos Tversky and Daniel Kahneman. “Judgment under uncertainty: Heuristics and biases”. In: *science* 185.4157 (1974), pp. 1124–1131.

[28] Alan Yuille and Daniel Kersten. “Vision as Bayesian inference: analysis by synthesis?” In: *Trends in cognitive sciences* 10.7 (2006), pp. 301–308.

[29] Carlos Zednik and Frank Jäkel. “Bayesian reverse-engineering considered as a research strategy for cognitive science”. In: *Synthese* 193.12 (2016), pp. 3951–3985.