Bachelor thesis

Artificial Intelligence

Radboud University 

The effects of irrationality on scientific agents in the simulation of scientific discovery

---

Author:
Benjamin Robijn
s-4822331
BenjaminRobijn@gmail.com

First supervisor:
dr. M. Blokpoel
Donders Institute for Brain, Cognition and Behaviour
m.blokpoel@donders.ru.nl

Second supervisor:
dr. I.J.E.I. van Rooij
Donders Institute for Brain, Cognition and Behaviour
i.vanrooij@donders.ru.nl

February 18, 2022

Abstract

The Rich et al. paper of scientific intractability [1] concludes that the solution to scientific inference is intractable given the current set of assumptions. This renders solutions to the problem of scientific automation unsolvable in polynomial time.

In an effort to further explore the idea of scientific automation derived from the notions of Rich et al. we attempt to use agent-based modelling to simulate the effects of irrational decision making on scientific agents. In doing so we explore the effects that irrationality has on agent's performance when solving scientific problems.

Agents representing scientists must ascertain the structures of a series of Finite State Transducers representing some cognitive system in reality that the scientists wish to understand. These scientific agents build up their own FST, their own theory of how the cognitive system works, incrementally as to try and match the structure of the cognitive system. By assigning these agents some level of irrationality which makes agents's FST building increasingly random as it increases we test by means of various metrics how the agents's performance changes for various levels of irrationality.

Ultimately, the findings from this paper indicate that adding irrationality to the agents's behaviour provides no added benefit to the effectiveness of the agents. An agent that chooses optimally every time will generally outperform an agent that makes sub-optimal moves for any metric this paper tests for. The agents perform better or on chance level for low irrationality and on or below chance level as irrationality increases. In the end, these findings remain in line with the Rich et al. paper of scientific intractability when it comes to the difficulty of automating scientific solutions.

Contents

Chapter 1

Introduction

1.1 Metascience: how to understand science

Metascience is the scientific field that is interested in the very concept of science itself. The ideas behind this field often lie in how science is most effectively conducted. Papers such as those of Smaldino et al. [3] and Stewart et al. [5] discuss the proper way of practicing science, often in relation to the reproducibility of results acquired. The field of metascience helps in understanding not only how science is performed effectively, but why certain things must be done in a certain way to allow scientific fields to grow properly. The 'reproducibility crisis' that the scientific community currently faces [2] is a good example of the importance of metascience in good science. The impracticality of reproducing results and the over-eagerness to get strong results, a behaviour this author has admittedly not been entirely innocent of in the past either, shows why we must keep strong focus on the proper conduct of science and maintain our understanding of the scientific process.

In context of the scientific process, one aspect explored in metascience is the process of scientific discovery.[4] The way humans develop scientific theories and iterate on them has proven to be an interesting topic of discussion with many applications outside of metascience itself. These applications range from uses in, for example, the fields of cognitive science and computer science, from being able to understand the psychological processes that bring about scientific discovery to being able to implement these processes into computerized form.

In this paper we take a look at one prominent example of the merging of cognitive- and computer science; that being the concept of automated scientists. For well over a decade the idea of creating smart scientific assistants, machines which can help humans in solving the various scientific problems

we face as a society, has been an alluring prospect and research in creating these scientists is still ongoing. [11]

And in taking interest in the notion of scientific automation we make special note of one paper, that being the paper of Rich et al.[1] which combines the psychological aspects of problem solving with the complexity theory often associated with computer science.

1.2 Rich et al. and the problem of cognitive intractability

The Rich et al. paper presents at heart a problem of computational intractability in a cognitive setting. The paper proposes a setting with a scientific agent and some cognitive system, meant to simulate a scientist in the real world attempting to understand the workings of this system. This cognitive system displays some situation-behaviour patterns, when some situation is presented, some behaviour comes out of the cognitive system which is observable by the scientific agent. An overview of this system is provided below in figure 1.1. The task for the scientists is to understand the structure of the cognitive system; what happens in this system that one specific situation goes in and one specific behaviour comes out? How can the agent ascertain the structure of this system without being able to observe it directly? This system of inference forms the foundation of scientific discovery as used in this paper.

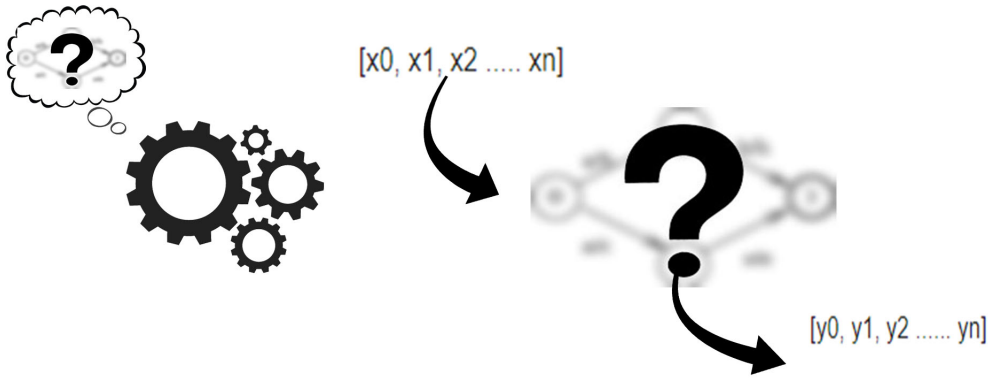


Figure 1.1: Scientist (Left) deliberating on the situation-behaviour pairs of a cognitive system (Right) in an attempt to ascertain the inner workings of the structure

In their paper, Rich et al. use a computation level explanation[1] to make clear the intractability of the inference problem presented to these scientific agents. Even given perfect data[2] the solution to acquire a proper theory for the agent is unsolvable in polynomial time.

This notion of scientific intractability and the process of scientific automation as used by Rich et al. serves as the foundation of this paper. It is important to note that this paper does not wish to challenge the notions made by Rich et al. and it is taken as truth that the process of scientific automation is intractable in this paper. Having stated as much, the focus will instead be placed on further exploration of the concept of scientific automation, with Rich et al. serving as the theoretical basis.

1.3 Exploring the automation of science

With the assumption of intractability established the focus must shift not to more efficient means of scientific automation[3], but rather the exploration of more effective scientific agents within that intractable space. Even if agents cannot be efficient, there is no reason that one agent cannot produce better results in scientific discovery than another, lest they all produce the same results for any system.

Instead, agents could be designed to behave very differently, and it is expected that they at least differ somewhat in their effectiveness for any given task. As such, to explore within the intractable space of scientific automation we take the same toybox reality Rich et al. [1] use in their paper and use it to test what different kinds of scientific agents could do differently with these cognitive systems.

An immediate line of reasoning to making changes to these agents would be to add some features that are biologically motivated. After all, when simulating real-life scientists, is it not reasonable to attempt to simulate with them some type of behaviour that may contribute in part to making this type of scientific inquiry possible in humans?

When thinking of these biologically motivated changes, one can simply look at the fundamental differences between man and machine for inspiration. And it is in that line of reasoning that the concept of *irrationality* is born. Machines are not perfect, but one would be hard pressed to call them irra-

[1] For the full computational proof, please see the paper attached in the bibliography

[2] No uncertainty, no faulty data, which by itself is unlikely in a realistic scientific setting

[3] The very notion of intractability eliminates any hope of 'efficient' agents, the two terms cannot coexist

tional, for their entire existence is based on logical systems. This then begs the question: would machines benefit from some level of irrational behaviour when being automated to perform scientific inquiry?

This in turn brings up the important main line of questioning this paper wishes to address.

1.4 Research question

"Do scientific agents with higher levels of irrational decision making perform better at solving scientific problems?"

1.4.1 Intuition

Testing out irrationality on the scientific agents described by Rich et al. is not simply about finding something to change and seeing what happens, there is a scientific intuition behind this line of questioning. Human irrationality can affect the way science is performed.[7] To put it in an alternative light: by deviating from what is considered the optimal solution for a scientific problem a scientist can potentially cause a benefit or detriment to the solution they may find. What if the scientific community attempts to solve a problem by means that may seem optimal in the short run, but ultimately wouldn't lead to a proper solution? By being irrational and sometimes making a sub-optimal choice we open up the possibility for alternative solutions in the search space of all possible solutions. Doing so may lead to novel ideas that may advance the scientific agents further towards a better solution that may otherwise not have been found.

This idea strongly correlates with the concept of local maxima, where perhaps a solution may be optimal in the local search space, but a much better solution can be found when looking outside of what is locally known by the agents. The idea behind irrationality in this context is for agents to be able to shoot themselves out of this local maximum of scientific reasoning so that they may reach the global maximum instead.

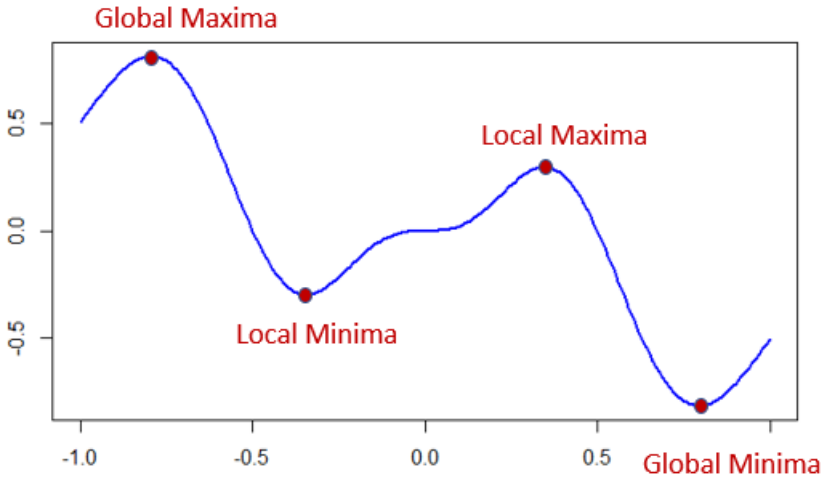


Figure 1.2: A simple illustration showing the concept of local and global maxima, showing that even if a solution seems good locally, it is not necessarily the best possible solution

1.5 Overview

Having covered the theoretical foundation in chapter 1, the coming chapters must shed more light on the notion of irrationality in science and a way must be devised to answer the research question as introduced in section 1.4. As such, this paper is laid out as follows:

To start, chapter 2 will go over two papers that concern themselves in some way with irrationality in a scientific setting. These papers will show irrationality in a broader scientific light and a quick comparison will be made between these papers and this one in order to provide motivation and context for the current research question.

Then, chapter 3 will go over the experimental design that encompasses specifications, assumptions, limitations, testing and consequent results of a model that is capable of simulating scientific discovery as described in the Rich et al. paper. [1] and chapter 1. This model should provide insights into the effects of irrationality on scientific agents and provide a basis on which to make interpretative remarks.

Finally, chapter 4 will provide analysis of the results gathered in chapter 3 and provide insights into what these results could mean in the broader context of cognitive automation and how it relates back to the original Rich et al. paper of cognition. Additionally, potential for future research will be

discussed as well as a clear conclusion to finally answer the research question in earnest.

Chapter 2

Related works

In order to establish an appropriate level of context for the concept of scientific irrationality introduced in chapter 1 the following chapter will go over two papers relating to the theme of irrationality as explored in the larger scientific community. These papers not only show a clear and present history of the scientific community's interest in this topic but will also hopefully convince any reader of the potential benefits irrationality might provide in scientific decision making.

The first paper by Julmi C. discusses the effects of intuition, a concept within decision making theory, and tries to establish a foothold in defense of intuitive decision making versus analytical decision making. The second paper by Chan L. et al. concerns itself with human irrationality in the context of a model of reward inference. In their paper they discuss the potential advantages and disadvantages of irrationality in agent-based models and make a general model for interpreting irrational decisions for agents to use.

2.1 The Julmi analysis of intuition effectiveness

2.1.1 Contents of the literature

In their paper Julmi. [9] has written a critical analysis of the concept of intuition. Intuition is a form of non-conscious, rapid decision making that lacks any sort of analytical thinking , only using information that is inherently present within the subject. What this means is that intuitive decision making is not a conscious process that is likely to be heavily influenced by an individual's biases and own ways of thinking.

Julmi goes on to explain that while intuition had in the popular zeitgeist been considered no more than an inferior form of decision making, not fed by logical decision but by the mere whims of the subconscious mind, there is now some precedent to think of it as something much different than that.

The analysis goes on to cite various pieces of research that have been undertaken to study the decision making process in humans. These studies have shown that while analytical thinking was thought to be the dominant stream of decision making with intuition merely being a 'shortcut' the mind would sometimes take to get to a fast answer, the two streams should actually be considered independant and not be considered interchangeable in the wider process of decision making.

Intuition, as is explained, offers better performance in decision making given certain parameters. Julmi explains that while often decision theory goes by the normative standard [1], humans actually posses unstable and often ambiguous preferences when making decisions. Intuition, as it turns out, is stronger when the 'structuredness' of a problem is low. What this means is that when a problem relies heavily on ambiguous information, that which the analytical stream cannot make sense of, the intuitive thought stream often has a better chance of solving the problems presented. Intuition also shows better results when it comes to equivocality, information with a high level of interpretability, as the snap-judgement characterized by intuitive thinking often leads to higher accuracies than when a subject attempts to analytically derive meaning themselves.

Julmi concludes that while these findings show clear benefits to intuition when decisions are made, no strong research has taken place yet to fully study the effects that this path of thinking may have, and encourages others to undertake research in a similar vein.

2.1.2 Remarks and relevance

Julmi makes a good case for intuition in humans being beneficial for various types of problems Of course, intuition as defined in the analysis is not directly irrationality as described in chapter 1 of this paper. Julmi says this as much in their paper, where intuition needs to be considered a separate form of rapid non-analytical decision making that *still adheres to rational thinking*. Intuition is inherently linked to a level of bias from the subject but still adheres to the rational that analytical thinking does, where it is coated in layer of learned habits or not.

While the definitions of intuition as presented in the analysis do not necessarily blend well with the definitions of irrationality we present in this paper, it *does* in abstract represent the same concept of alternate decision making we aim for. By exploring the notion that various other forms of decision making can yield some benefits compared to completely optimal

[1] The expectation that all decisions are born from rational thinking

decision making, we can establish that there is the potential irrationality as described in this paper may yield some benefit too.

2.2 The Chan et al. paper on the advantages and disadvantages of irrationality for reward inference

2.2.1 Contents of the literature

In reward inference an agent must observe behaviour as generally displayed by humans and derive a reward function from that behaviour so that we may explain how well certain behaviours lead to certain rewards. As Chan et al. [6] note this inference problem goes by the assumption that humans behave rationally when working towards a goal, yet humans themselves tend to display various forms of irrational behaviours that to a machine would be sub-optimal.

As such, Chan et al. strive in their paper to model how these various irrationalities affect the overall process of reward inference. In order to do this they use the bellman optimality equation often used for reinforcement learning and design it in such a way so to represent various types of irrational behaviour. For example, negative or positive bias values can be assigned to the function in order to simulate positive or negative bias of an individual. In doing so, they could emulate the irrational behaviour of humans in analytical form and compare this model to purely rational inference functions. [2]

When their model and testing was finished Chan et al. took note of a few interesting findings. First of all, modelling irrationality could outperform rationality when modelled correctly. It was of importance that the agent was modelled as irrational rather than noisily-rational, failing to correctly account for the irrationality aspect in question lead to the agent performing far worse than even just taking prior beliefs and skipping inference all together. It is noted however that there is only need to *approximate* the irrationality aspect. Should for example the system interpret irrational behaviour in the form of myopia [3] but get the type of myopia wrong, the system would still perform better than a system that was wholly rational. It was found that irrationally was just generally more informative than rational behaviour, as irrationalities allowed for a strong change in reward values, so was it easier for the agents to infer the reward function.

[2] For full explanation of the methods used, please see the Chan et al. paper in the bibliography

[3] Nearsightedness, in the context of the model failing to account for long term effects of actions

In conclusion, Chan et al. found that while modelling irrational behaviour as merely noisily-rational was a great detriment to to the information gain of the reward inference, even correctly approximating the irrational behaviour lead to greater successes than modelling the agents as rational. They remark that modelling irrational behaviour in agents could prove to be beneficial to their performance and that more research is needed into this topic for far more diverse areas in order to properly understand all the applications this may have.

2.2.2 Remarks and relevance

As convincing arguments go, the Chan et al. paper provides great examples in favor of the research subjects of this paper. Modelling irrationalities and finding that they proof beneficial to performance of agents is a great foundation on which to build this paper and see if the model as presented in chapter 3 can show the same results as the Chan et al. paper or if something is instead different in the findings.

It should be noted as an aside that Chan et al do consider various forms of irrationality in their model while this model just considers a general 'irrationality factor'. Whether or not this generalization may be too simple to accurately model irrationality will have to be shown in the coming chapters.

Chapter 3

Experimental design

Having established the theoretical foundation of this paper in chapter 1 and given context for the research subject in chapter 2, we must now find a way to answer the research question as proposed in section 1.4.

In order to do so and as such test irrationality for scientific agents in a simulation of scientific discovery we take inspiration from the paper of Rich et al.' [1] and use the Scientific agent-cognitive system concept as used there. We take an agent that has to infer on the structure of some cognitive system by creating his own theory, with the aim of that theory being to match the cognitive system as closely as possible.

3.1 The cognitive system interpretation model

3.1.1 General functionality of the model

The main purpose of this model is for a scientific agent to be able to display irrational behaviour while trying to match the structure of some cognitive system. In order to do this we must first define a few concept within that model before we can give a specification of how it is supposed to work.

The structure of the cognitive system is by itself unknowable to the scientific agent, for it can only see the situation-behaviour pairs as explained in section 1.2. This necessitates the model to incorporate a cognitive system that can take in some form of data and output it in the same way so that the agent may easily interpret the results.

For this purpose we have taken a Finite State Transducer, hereby known as FST, as the basis of the cognitive system. FSTs are networks consisting of one or more states that can be traversed through by means of transitions,

an example of a simple FST can be found in figure 3.1 . When some input is given to an FST, it takes the first symbol [1] of that input and attempts to transform it to some output symbol. For this to happen, the state the machine is currently on needs to have a suitable transition present. Given the example of figure 3.1, where a B leads from s0 to s0 and transforms into an A and vice versa towards s1, should s1 be reached the system can no longer transform any more data as no outgoing transitions are present. Should no transitions be present to move the input of the FST forward, the FST returns the output it currently has and stops.

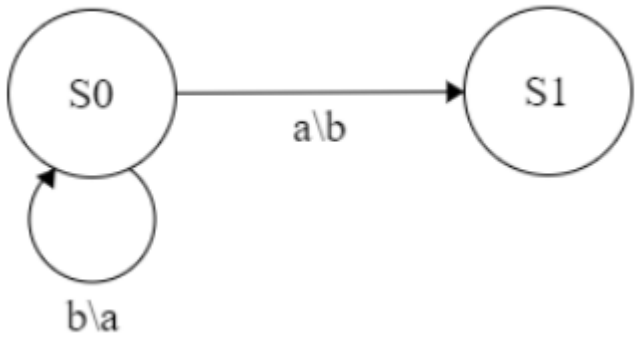


Figure 3.1: A simple example of a Finite state transducer as representing the cognitive system, with its states (circles) and transitions (arrows)

FSTs are a strong means of representing scientific theorems as they are easily manipulable, interpretable, and as such make for non-ambiguous means of interpretation of the effectiveness of the agent's performance.

Given some FST as representing the cognitive system, the scientific agent must match the structure of this system well in order to succeed at its theory building. However, without being able to see any structural information of the cognitive system the agent must rely on the situation-behaviour pairs as displayed by the unknown FST. Assuming that the agent has perfect rationality and always makes the best move, the agent simply takes its own theory FST, checks which mutation [2] causes the output of its own theory to match the cognitive system's best [3], and then makes that change.

This model doesn't always allow the agents to be rational however. Every

[1] Symbols can mean anything that functions as input such as numbers or variables, but are within this model always individual letters

[2] For an FST in this model, a 'mutation' is the adding of a transition to the existing theory or adding a state and making a transition towards it

[3] This is defined as the 'fit' on data, as used below

agent gets assigned some value of irrationality between 0 and 1. Whenever the agent checks all possible mutations as explained above, instead of merely going for the mutation that fits the data best, it puts the % match between the data of agent and cognitive system for every mutation in a list and uses the softmax function to simulate irrational behaviour.

$$\text{softmax}(x)_i = \frac{e^{\frac{y_i}{T}}}{\Sigma_j^N e^{\frac{y_j}{T}}}$$

In this function the temperature as denoted by T allows irrational behaviour to occur. When T is low, the agent tends to pick the mutation that has the best % match between the mutation FST and the cognitive system. However, as T increases, and thus the irrationality factor, the mutations list becomes more and more equal in probability to be chosen, this leads to the agent making sub-optimal choices more often, simulating irrational behaviour.

The eventual goal for the agent is to match the situation-behaviour pairs of the cognitive system to a sufficient level. This is indicated by some accuracy threshold, when the two pairs match well enough, the agent stops changing its theory and as satisfies with what it has created.

3.1.2 Model specification

Below is given the model specification as to help in illustrating the behaviour as explained in section 3.1.1:

Cognitive system

A cognitive system M is represented by a (F)inite (S)tate (T)ransducer. This is a sex-tuple $(\Sigma, \Gamma, Q, q_0, \delta, \omega)$ where:

$$\begin{aligned}
\Sigma &= \text{a finite non-empty set of input symbols} \\
\Gamma &= \text{a finite non-empty set of output symbols} \\
Q &= \text{finite set of states} \\
q_0 &= \text{initial state, } \in Q \\
\delta &= \text{state transition function: } \delta : Q \times \Sigma \to Q \\
\omega &= \text{output function: } \omega : Q \times \Sigma \to \Gamma
\end{aligned}$$

Where M receives certain situations S where $S \in \Sigma$ and outputs behaviours B where $B \in \Gamma$. This pairing forms data set D where $D : S \times$B

Scientific agent

A scientific agent A is a quadruple (M', θ, α, I) where:

$$\begin{aligned} M' &= \text{FST representing agent's own theory} \sim M \\ \theta &= \text{acceptance threshold: } 0 \leq \theta \leq 1 \\ \alpha &= \text{complexity limit: } \alpha \geq 0 \\ I &= \text{an irrationality factor: } 0 \leq \mathrm{I} \leq 1 \end{aligned}$$

Where A must build up M' to match M based on I, $(D \in M)$ and $(D \in M')$, with θ and α as stopping conditions.

Input:
A receives a data set D based on the situation-behaviour pairings it can observe from M. Where each $(S\ ,\ B) \in D$ consists of a string of symbols as contained by Σ and Γ.
A also receives a previously hypothesised system M', which may be an FST consisting of zero states and transitions (e.g. when the agent wishes to build a new theory) or already contain a theory based on the agent's previous iterations.

Output:
A creates M'' by making single mutation to M'. This mutation can either be a transition to an existing state given that the same state does not have two outgoing transitions with the same input symbol, or the creation of a new state and a transition to that state. All possible mutations for the current M' are stored in list $L : l \in L$, where l represents the fit of a mutation's situation-behaviour pairs to the cognitive system and $0 \leq l \leq 1$. This fit of data is achieved by comparing the input-output pairs of l and M pairwise, meaning every symbol is matched to every symbol. Should the length of one output be different from the other, then all excess symbols are counted as not-matching.

Given L, I is used to transform L by means of the softmax transformation function into a probability distribution where $\mathrm{sum}(L) = 1$ and L becomes more uniform as I increases and more skewed towards $\max(L)$ as it decreases. Based on this transformed probability distribution a mutation is chosen based on the probabilistic chance that L provides the agent. The fit on data of the chosen mutation is hereby known as Y and M' with the new

mutation is known as M'' .

M'' is then checked by using the stopping conditions. If $Y > \theta$ or the number of states in $M'' > \alpha$ the agent stops looking for a better theory and outputs M'' alongside Y as it is satisfied with the theory it has created. If neither of these conditions are met, M'' becomes the new M' and the process begins anew.

3.1.3 Implementational Assumptions

The model's functionality as described in detail in sections 3.1.1 and 3.1.2 works fine on a theoretical basis but when it comes to putting it in practice some assumptions and limitations do come to light. The following sections shed light on some of these findings and in what way they influence the implementation of the theoretical model.

First of all, in the theoretical model any state within the FST structure could be a starting state. while this could still hold for the implementation part of the model it was deemed more reasonable for manual testing of results to always have the starting points be assigned a fixed point, rather than jumping randomly between iterations of the model. As such, the starting state has now been fixed at s0, making it easier to interpret behaviour and results when manually checking the process.

Additionally, both the amount of input-output pairs as well as the length of input and output are of arbitrary length in the theoretical model. Various boundaries had to be set in the implementation of the model as to strike a balance between informative runs of the model versus amount of testing being possible given computational resources. In the end. The amount of symbols in a single string of input has been set to 25, with every data data set containing 50 pairs of situation-behaviour. After many trials this amount of data was deemed ideal in balancing the needs of testing without making the model not informative enough. Setting these values too high made effective testing impossible while setting them too low caused the agents to be satisfied very quickly as the little amount of data processed often led to high fits on the data due to chance. In a similar vein, the symbols list used as the input and output symbols of the model was set to 6 [4]. This was done for the same reasons as the input and output pairings.

In the context of the symbols, an assumption was made that the data presented to the agent could never be made of symbols not listed within the input and output lists and that these list were the same. In the theoretical

[4]More precisely, ['a', 'b', 'c', 'd', 'e', 'f']

model both the input and output symbols lists could differ, though it was deemed unnecessary in the implementation due to how the model works. As it stands, the model is capable of handling unknown symbols in that it would merely return the output string as is when a proper transition is not presented to it within the state it is in. As such, adding symbols that are not represented in the FST structures would merely be adding random chance that output was cut short sooner than necessary, which was deemed unnecessary.

Finally, the FSTs in this model are through their process of creation made to be deterministic. What this means is that a state can have no more than one transition with the same input symbol. The reasoning behind this assumption is twofold: One, it reduces the random chance factor in determining situation-behaviour pairs, making so that whenever a letter is considered in a state the agent does not have to flip a coin on which transition he moves. Two, it makes so that whenever the agent needs to consider all possible mutations to his theory, it significantly reduces the workload for the agent, which helps in reducing the exploding search space problem; a problem that is very clearly present in the model.

3.1.4 Limitations

As generally hinted in throughout this paper there also a few limitations that this model runs into, this is both in the theoretical model stage as well as in the implementational stage. These limitations are as follows:

The primary limitation this model runs into is that is *very* difficult to ascertain structural information purely from the fit on the data presented. An agent can generally form a bias using fit on data towards getting the structure more right than chance level agents, but as Rich et al. [1] also indicate, finding a proper solution towards matching structural information is more difficult, so while interesting on a theoretical basis, a good structural fit seems unlikely.

When it comes to building theories, the agent cannot remove any edges from its theory. This means that once an edge has been chosen it cannot be changed. This choice was made to simplify the building of theories and to prevent the agent from getting stuck in a loop. However, given the sometimes arbitrary edge choices in the beginning of the process it is very likely that the agent cannot match the structure of the cognitive system perfectly as this would require it to get every edge right from the start. This being extremely unlikely means that the system leans more towards lower structure similarities from the start.

With the way cognitive system FSTs are randomly created it is also possible for the cognitive system to have a 'state island'. This is when two or more states are connected to each other but not to the rest of the FST. While it is possible for the cognitive system to have this structure, with the incremental building of the agent's theory this behaviour is not replicable. While these state island are relatively rare, whenever one does appear it becomes impossible for agents to exactly match that structure, further lowering structural match averages.

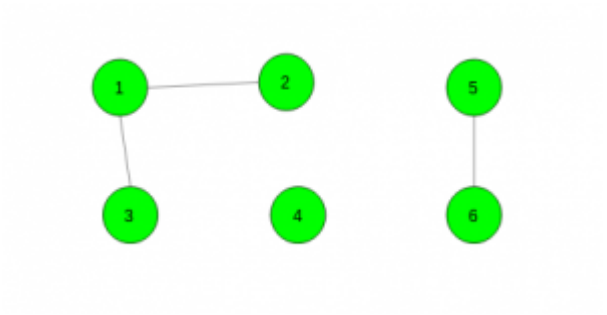


Figure 3.2: The possible isolation of nodes in an FST, this behaviour is unreplicable for agents when incrementally building their own theories

Finally, this model does not take into account graph isomorphism. [5] With the way structure similarity between two FSTs is tested there is no way for the model to understand that two FSTs which are similar in structure but are, for example, inverted are more alike than two FSTs where only the first transition matches between the two. While the incorporation of a proper solution would have been preferable, it was deemed ill-advised to account for this limitation given the relative small impact isomorphism actually has on agent's accuracies and the notorious difficulty this issue presents.

3.2 The testing process

With all information about the model and how agents are to build their theories established, it is now time to actually test the effects that irrationality might have on scientific agents. One question that can easily be asked of this is what exactly to test for and to consider when assessing performance of these agents. Given that structure similarity between agent and cognitive

[5] The notion that two graphs are similar due to the same structural elements but do not necessarily appear that way

system may be low given the fit on data as explained in section 3.1.4, can you really base the effects irrationality has on that metric alone?

3.2.1 General outline of testing

Metrics used

Given that structure similarity between the agent's theory and cognitive system is the most important metric of testing but perhaps not the most informative by itself, it was instead decided to use a group of metrics to test the general performance of agents when building scientific theories. The intuition being that while fit on data or structure similarity may not tell the whole story by itself, a combination of relevant metrics could indicate the general effects irrationality has and build confidence in those findings.

In the end, it was decided that the following metrics would be used to support the research question's answer: The fit on data between the agent's theory and the cognitive system, the % match in size of the two FSTs, the number of transitions the agent's FST has and finally the structural similarity between the agent and cognitive machine.

Structural similarity would be the most prominent metric, it most clearly illustrating whether or not the agents managed to correctly predict the cognitive system's structure. Match in size is an approximation of the match in structure, with the hope that even if match in structure proves to be uninformative, that the match in size might still be show some interesting results.

The fit on data is a general baseline of how irrationality affects the agents. It is expected because the agents are trained on data that this metric would perform well, but it might still show a general effect amongst the irrationality levels that can support the findings the other metrics produce.

Finally, the number of transitions is used to determine how fast an agent can produce results. When measured in combination with the match in structure size this could prove interesting. For example, when an agent can find a solution with fewer transitions but manages to get a lower rating on the size match, that would actually be a detrimental effect instead of a beneficial one.

Specifics of testing

The four metrics as mentioned above where tested for 80 agent with varied levels of irrationality factors. Each agent was assigned a random data pool

of 50 data points with up to 25 symbols [6] as described in the assumptions. The complexity limit was set to 7 states and the accuracy threshold set to 75% match on data. This accuracy threshold came about when testing what threshold would still allow for complex FSTs but would not be too hard for agents to match and prove uninformative. In order to generalize findings each agent was trained on the same 50 cognitive systems. The results as shown below show these generalized results, alongside error bars and the chance level results. The chance levels were acquired by creating two random FSTs and checking them for the same metrics as the agents.

3.3 The Results

3.3.1 Agent theory vs cognitive system, fit on data

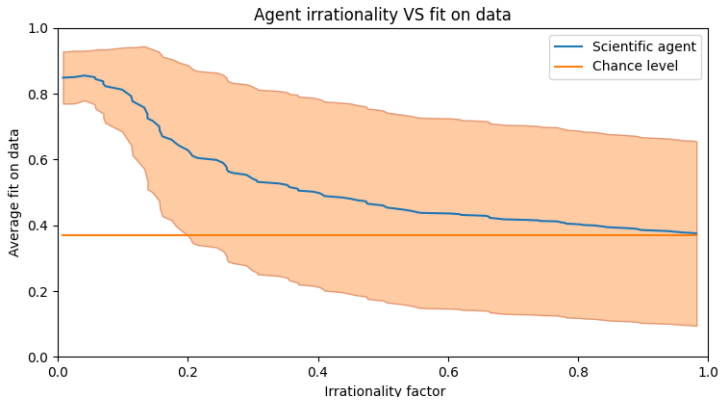


Figure 3.3: Fit on data

The expectation that the fit on data metric would do well has been somewhat validated by these results. Agents with low irrationality factors perform very well when matching the data between the cognitive system and the agent's own theory. These percentages peaking at 90% match do however quickly lower towards 50% as irrationality increases. When the agent is maximally irrational it perform no better than chance level. Of note is that while generally these agents perform well in fitting the data, there are still wide margins of error, with error bars increasing to about 25% as irrationality grows.

[6] but no less than 3

3.3.2 Agent theory vs cognitive system, number of transitions

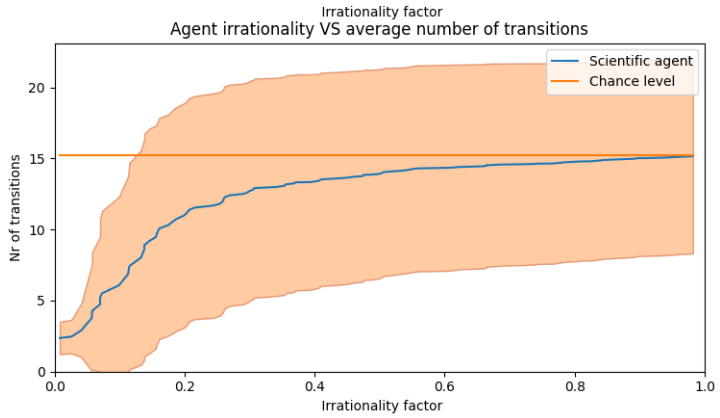


Figure 3.4: Number of transitions

The number of transitions metric shows a clear increase in transitions as the irrationality factor increases. The number of transitions are low for lower irrationality levels but likewise increases as irrational behaviour becomes more prominent until reaching chance level at maximum irrationality. Error bars show some level of variance in the FSTs but the general line remains that as the agents become more irrational they need to use more transitions in order to get to a theory that is satisfactory. This indicates some level of 'fumbling' for these agents, as the individual information gain from a transition is lower, which is in line with the general strategy of sub-optimal' movement.

3.3.3 Agent theory vs cognitive system, % match in size

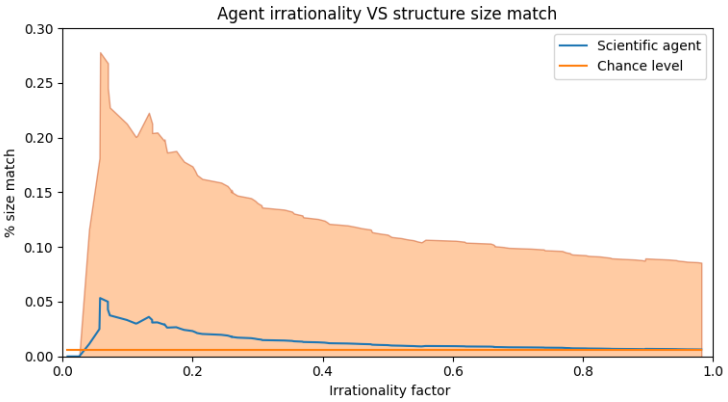


Figure 3.5: % Structure size

The % match in structure size shows a curious trend. More often than not[7] the accuracy in structure size match peaks not for maximum or minimum irrationality, but for slightly less than minimal irrationality. The agent performs better at lower irrationality levels overall still but performs worse when there is no irrationality at all. The leading theory for this is that agents that are perfectly rational are simply too good at what they do. With the number of transitions being low as seen in the number of transitions result it would seem that agents get to a satisfactory theory too quickly, but fail to match the complexity of the cognitive system they are trying to replicate. This in turn opens an interesting line of thinking where in order to capture all nuances of a concept we need to teach agents to not always perform optimally, lest they simplify a system that is simply too complex for that way of thinking. Of note is however that this trend is not present in all runs of the model, but most of them. That, combined with a relatively large margin of error might not allow this theory to be set in stone just yet.

3.3.4 Agent theory vs cognitive system, structural similarity

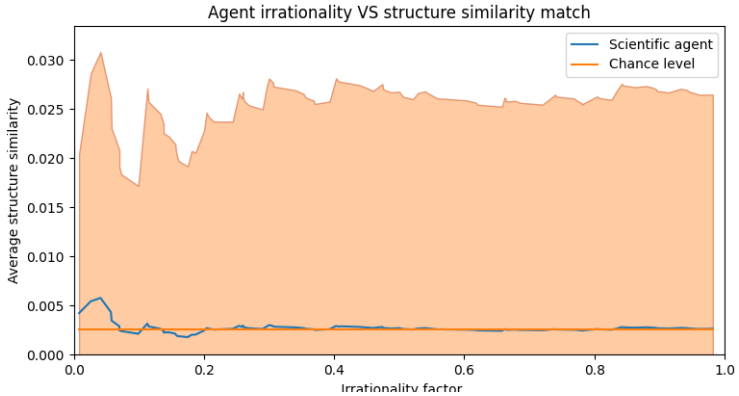


Figure 3.6: % Structure similarity

Structural similarity for agents is quite low, sitting at roughly 0.05% accuracy. While there is a consistent trend of optimal decision making being ever so slightly better than chance level in matching the structures of the cognitive systems, this effect is not strong given the large error margin and small accuracy to begin with. These findings seem to be in line with both the Rich et al. [1] paper of complexity as well as our admittance in limitations of the model. Simply put, while there is seemingly a small benefit to optimal decision making, this is so small a benefit that it can not be called a trend at all. This is even more evident in other trials of the model, where sometimes lower irrationalities would actually perform at or below chance

[7]For all results, please see the appendix

level. All in all, it would seem that irrational decision making has barely any effect on agents's performance compared to chance level.

3.3.5 General analysis of results

All metrics used in this paper except for the structural match showed some level of advantage over chance level for lower irrationality levels. As irrationality increased the performance of agents dropped to chance level, with chance level itself being reached at maximum irrationality. Lower irrationality preferred by the fit on data and number of transitions, but the number of transitions and structure size match made a compelling argument for some level of handicap for scientific agents as one may need to limit their capabilities in order to capture all the nuances of a problem. In the end, the most important metric of structural match between agent and cognitive system showed a simple trend. That being that high irrationality offered no better performance than chance level and lower irrationalities seemed to perform better worse, or at level at random.

Chapter 4

Discussion

4.1 Interpretation and implications of the findings

With the results found in chapter 3 one can finally attempt to answer the research question in earnest. *"Do scientific agents with higher levels of irrational decision making perform better at solving scientific problems?"* Within the parameters of this paper that answer would be a decisive no. While a small trend was found relating to scientific agents possibly being too good to capture all the nuances of a problem, potentially linking it to the Chan et al. [6] paper of advantages and disadvantages of irrationality in reward inference, this trend cannot be more than casually linked to their findings however, as the effect is simply too small to be strongly relevant. It would seem that irrational decision making as produced in this paper merely drives agents to chance level judgements when they become too irrational. It should be stated that these findings do align with the Rich et al. [1] paper of the complexity of automating science. It is therefore also perhaps not surprising that irrationality as modelled here has no clear effect on performance given the complexity of the task.

4.2 Limitations of the study and further recommendations

As mentioned in the limitations section of this paper, the introduced notion of irrationality as being merely the lack of optimal choices seems to be too simple a thought process for this kind of research. Chan et al. [6] has already shown that modelling irrational behaviour can be beneficial to the training of agents. This in turn leads this author to believe that while the notion of irrationality might be too simplified in this paper, the notion of irrational decision making being beneficial still stands. Taking inspiration from Chan et al. one could model various levels of irrational behaviours. For example, given the same task of matching a scientific theory, an agent could

focus on a specific characteristic of that problem. Perhaps they believe that any state could only have one transition? Perhaps it favors the creations of state island, believing the ideas as presented in the cognitive system to not be linked at all? They are but simple suggestions, but if Chan et al. has shown one thing its that the potential exists for human behaviour to influence agents in a beneficial way.

4.3 Conclusion

Following the conclusions from Rich et al. this paper attempted to show whether irrationality in scientific agents might be beneficial in the automation of scientific discovery. The results shown however do not make that effect likely. Still, with relevant and promising research taking place in the field of decision theory it might still be an idea working into. This paper concludes that while simulating irrationality in too simple a manner proves to be detrimental to agents, there could still be a future for human imperfections in machines given that they are modeled properly.

Bibliography

[1] Rich, P. (2021), 'How hard is cognitive science? Proceedings of the Annual Meeting of the Cognitive Science Society', Cognitive Science Society,
https://escholarship.org/uc/item/8cr8x1c4

[2] Piper, P. (2020), 'Science has been in a 'replication crisis' for a decade. Have we learned anything?'. Vox,
https://www.vox.com/future-perfect/21504366/science-replication-crisis-peer-review-statistics

[3] Smaldino, E.. (2016), 'The natural selection of bad science', The Royal Society Publishing,
https://royalsocietypublishing.org/doi/10.1098/rsos.160384

[4] Langley, P. (1995), 'Stages in the process of Scientific discovery', Institute for the Study of learning and expertise,
https://www.aaai.org/Papers/Symposia/Spring/1995/SS-95-03/SS95-03-019.pdf

[5] Stewart, A.J. (2021), 'The natural selection of good science', nature human behaviour,
https://www.nature.com/articles/s41562-021-01111-x

[6] Chan, L. (2021), 'Human irrationality: both bad and good for reward inference',
https://arxiv.org/pdf/2111.06956.pdf

[7] Potochnik, A.P. (2020), 'Awareness of our biases is essential to good science', Scientific America,
https://www.scientificamerican.com/article/awareness-of-our-biases-is-essential-to-good-science/

[8] Dilmegani, C.D. (2021), 'What is Cognitive RPA: in-Depth Guide to RPA's future in 2021', AiMultiple,
https://research.aimultiple.com/cognitive-automation/

[9] Julmi, C. (2019), When rational decision-making becomes irrational: a critical assessment and re-conceptualization of intuition effectiveness., Bus Res 12, 291–314, https://doi.org/10.1007/s40685-019-0096-4

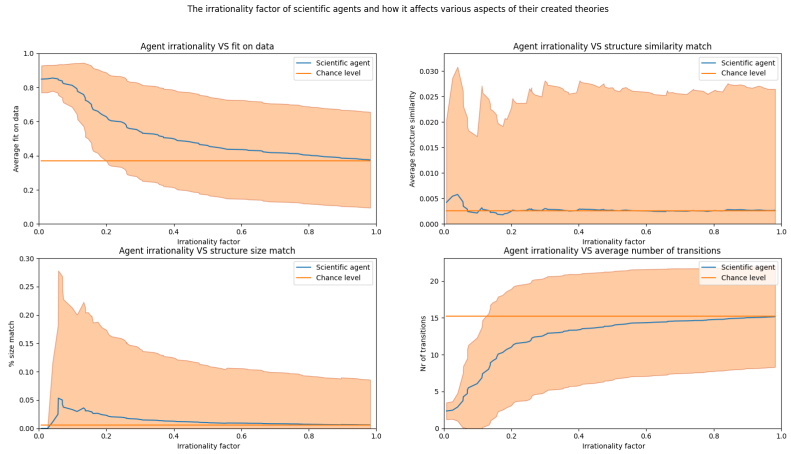
[10] Bonabeau, E.B. (2002), 'Agent-based modelling: Methods and techniques for simulating human systems', PNAS, https://www.pnas.org/content/99/suppl/7280

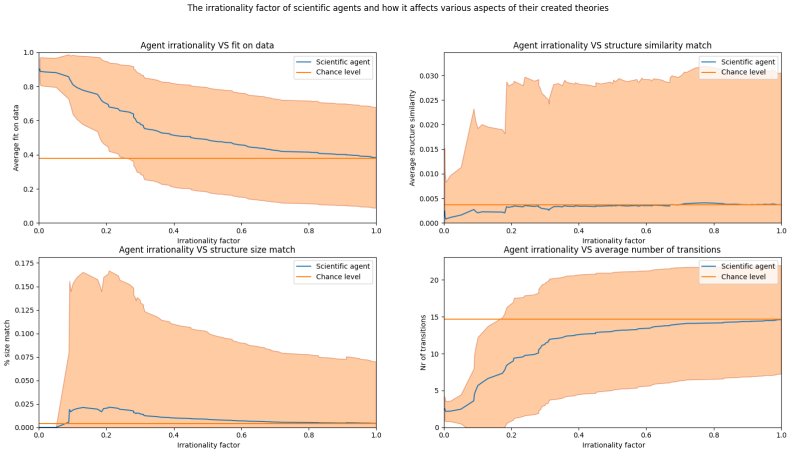
[11] Waltz, D. (2009), 'Automating science', Science, https://www.science.org/doi/full/10.1126/science.1172781

[12] Devezer B. (2019), 'Scientific discovery in a model-centric framework: Reproducibility, innovation and epistemic diversity', Journals, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0216125

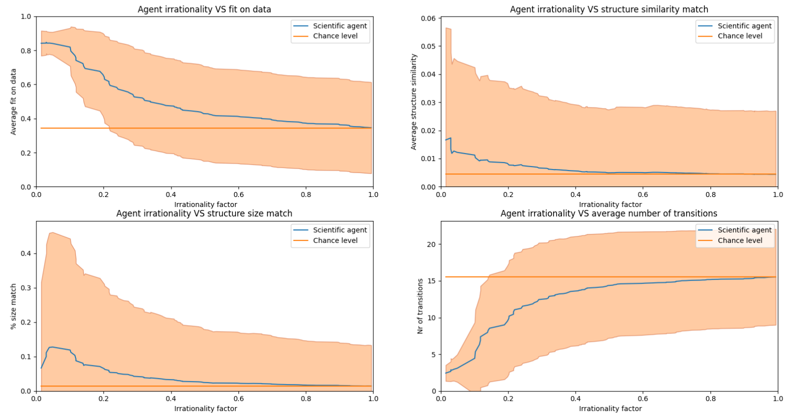
Appendix A

Result list

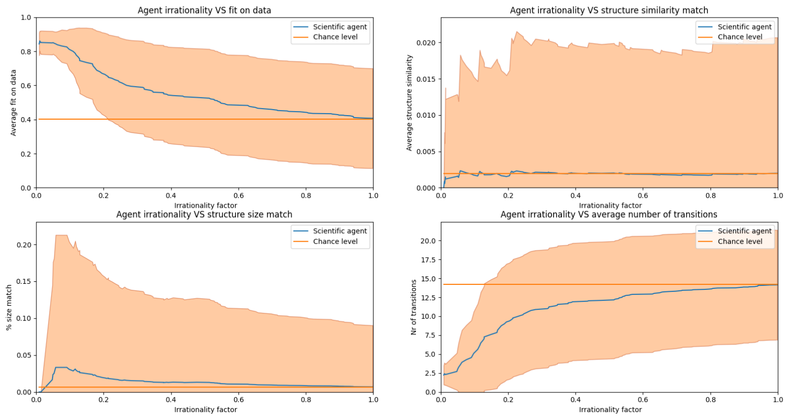




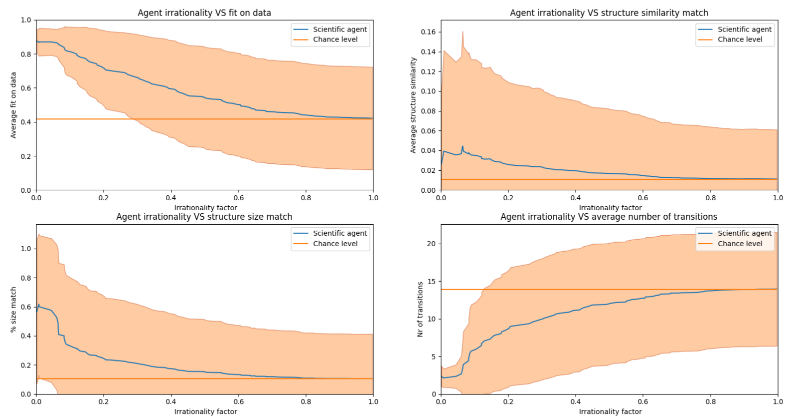
The irrationality factor of scientific agents and how it affects various aspects of their created theories

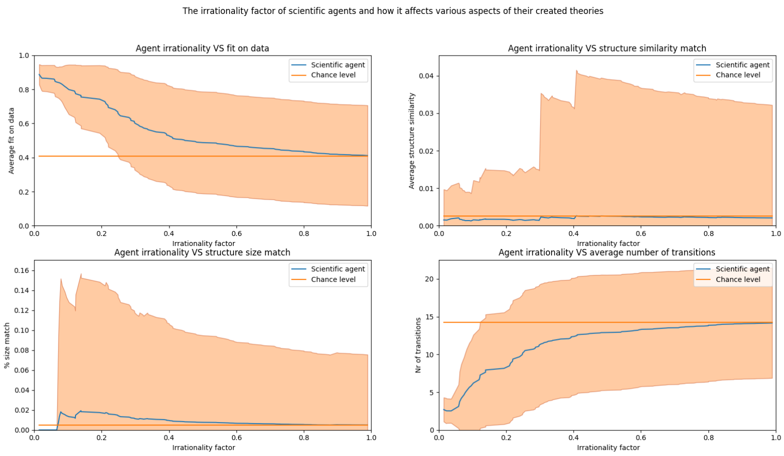


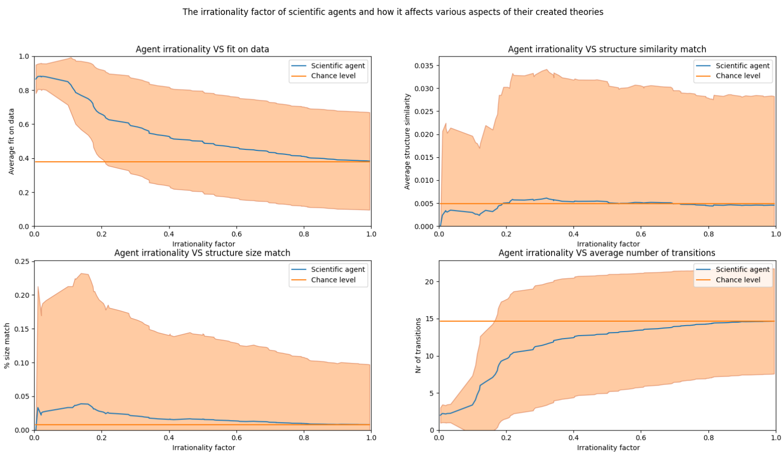
The irrationality factor of scientific agents and how it affects various aspects of their created theories

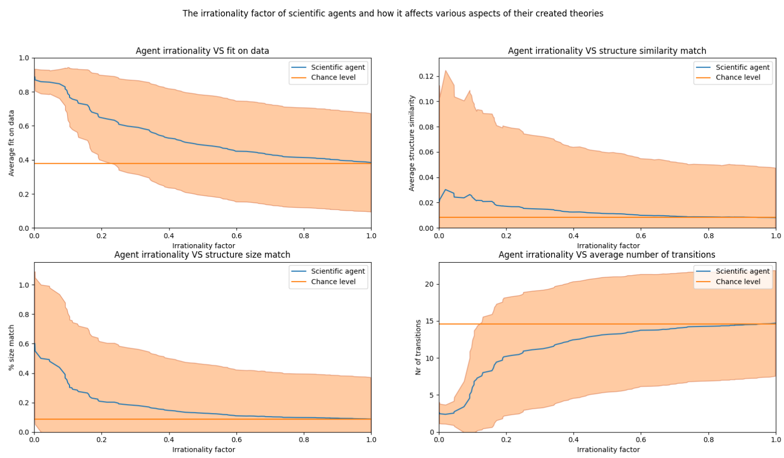


The irrationality factor of scientific agents and how it affects various aspects of their created theories

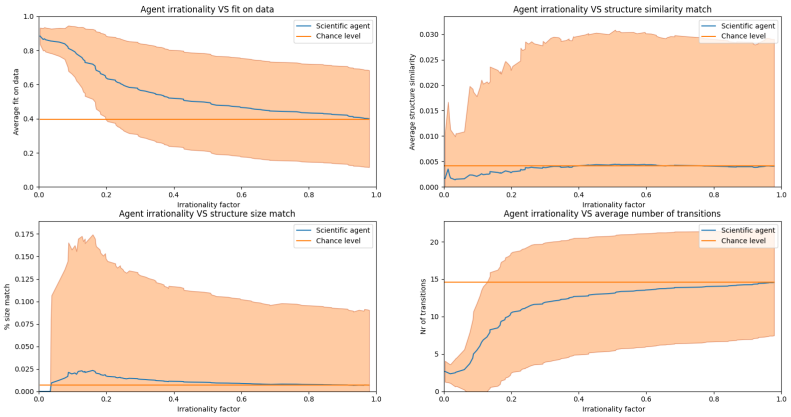



The irrationality factor of scientific agents and how it affects various aspects of their created theories
Agent irrationality VS fit on data
Average fit on data
Irrationality factor
Agent irrationality VS structure similarity match
Average structure similarity
Irrationality factor
Agent irrationality VS structure size match
% size match
Irrationality factor
Agent irrationality VS average number of transitions
Nr of transitions
Irrationality factor
Scientific agent
Chance level

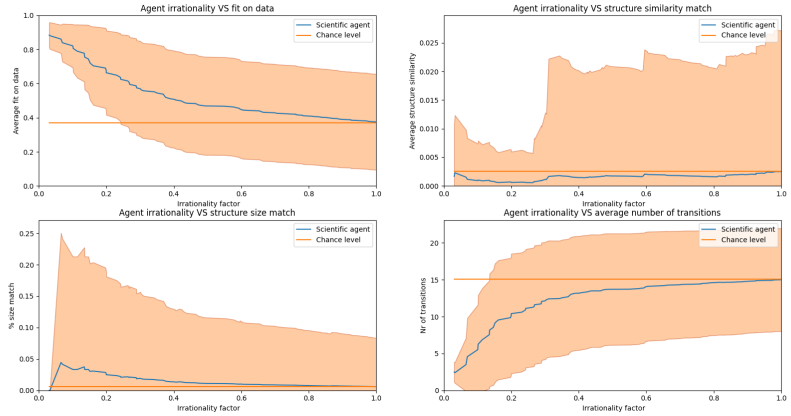

The irrationality factor of scientific agents and how it affects various aspects of their created theories
Agent irrationality VS fit on data
Average fit on data
Irrationality factor
Agent irrationality VS structure similarity match
Average structure similarity
Irrationality factor
Agent irrationality VS structure size match
% size match
Irrationality factor
Agent irrationality VS average number of transitions
Nr of transitions
Irrationality factor
Scientific agent
Chance level


The irrationality factor of scientific agents and how it affects various aspects of their created theories
Agent irrationality VS fit on data
Average fit on data
Irrationality factor
Agent irrationality VS structure similarity match
Average structure similarity
Irrationality factor
Agent irrationality VS structure size match
% size match
Irrationality factor
Agent irrationality VS average number of transitions
Nr of transitions
Irrationality factor
Scientific agent
Chance level

The irrationality factor of scientific agents and how it affects various aspects of their created theories



The irrationality factor of scientific agents and how it affects various aspects of their created theories



Appendix B

Source code

The repository found at:

https://github.com/Chibitasilver/thesis_code_simulating_irrationality_in_scientific_discovery

Contains the code for the model used in this paper alongside documentation and higher resolution versions of the plots found in the appendix and results sections.