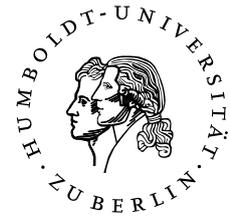


HUMBOLDT-UNIVERSITÄT ZU BERLIN
MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT
INSTITUT FÜR INFORMATIK



A Conversational Agent for the Improvement of Human Reasoning Skills

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)

eingereicht von: Laura Wartschinski

geboren am:

geboren in:

Gutachter/innen: Prof. Dr. Niels Pinkwart

Dr. Nguyen-Thinh Le

eingereicht am:

Contents

1	Introduction	5
2	Reasoning	7
2.1	Terms and Definitions	7
2.2	Heuristics and Biases	8
2.3	Dysrationalia	8
2.3.1	Lacking Mindware	9
2.3.2	The Dual Process Theory	9
2.4	Importance of Rational Decisions	10
2.5	Teaching Reasoning	10
2.5.1	Challenges	10
2.5.2	Prospects for Teaching	11
3	Software for Teaching Reasoning	12
3.1	Applicability to the Reasoning Subject	13
3.2	Features of Believable and Effective Pedagogical Agents	13
3.2.1	Effective teaching	14
3.2.2	Social Roles	14
3.2.3	Off-task Communication and Chatting	15
3.3	Computers as Social Agents	15
3.3.1	Personality of a Social Agent	16
3.3.2	The Use of Emotion	17
3.3.3	Encouragement and Praise	17
3.4	Animated and Text-Only Representations	18
3.5	Summary of Requirements	19
4	The Agent: Liza	19
4.1	Conceptual Design	20
4.1.1	Introduction and Greeting	20
4.1.2	General Process	21
4.1.3	Mixed-Initiative Dialogue	22
4.1.4	Asking Questions	24
4.1.5	Explaining	26
4.1.6	Feedback	27
4.2	Parsing	27
4.3	Finding Tags	28
4.3.1	General Parsing	28
4.3.2	Situation-Specific Parsing	28
4.4	Architecture	29
4.4.1	Control Unit	30
4.4.2	UI and GUI / Interface	31
4.4.3	Stories and Storystore	31

4.4.4	Phrasebase	31
4.4.5	Parser	32
4.5	Implementation	32
5	Teaching Content	33
5.1	Topics Taught by the Agent	34
5.1.1	Bayesian Reasoning	34
5.1.2	Law of Large Numbers / Regression to the mean	35
5.1.3	Gambler’s Fallacy	36
5.1.4	Wason’s Selection Task	37
5.1.5	Covariance Detection	38
5.1.6	Sunk Cost Fallacy	39
5.1.7	Belief Bias in Syllogistic Reasoning	40
5.2	Relation between the reasoning tasks	41
5.3	CRT	41
5.4	Critical Thinking and Argumentation Skills	42
6	Experimental Evaluation	43
6.1	Goals	43
6.2	Design and Setup	44
6.2.1	Origin of Texts and Tasks	46
6.3	Conduction of the Evaluation Study	47
6.3.1	Characteristics of the Participants	47
6.4	Results	49
6.4.1	Sunk Cost Fallacy	50
6.4.2	Gambler’s Fallacy	50
6.4.3	Bayesian Reasoning	51
6.4.4	Belief Bias in Syllogistic Reasoning	51
6.4.5	Regression towards the mean	51
6.4.6	Covariation Detection	51
6.4.7	Wason’s Selection Task	52
6.4.8	Influence of Age, Gender, Education and Language Proficiency	52
6.4.9	Cognitive Reflection Test	52
6.4.10	Personal Feedback by the Participants	52
6.5	Discussion	53
6.5.1	Confirmation and Rejection of Hypotheses	53
6.5.2	Possible Explanations	54
6.6	Limitations	54
6.6.1	Number and characteristics of participants	54
6.6.2	Previous Knowledge	55
6.6.3	Scope of Questions	55
6.6.4	Time Spent with the Agent	55
6.6.5	Long term effects	56
6.6.6	Minor Concerns	56

7 Conclusion	56
8 Future Directions	57
8.1 Improvement of the Agent	57
8.1.1 Refining the dialogue	57
8.1.2 Further Personalization of the Learning Plan	57
8.1.3 Scaling Up the Learning Experience	57
8.1.4 Expanding the Scope	58
8.2 Further Research	58
Appendices	66

1 Introduction

With the increase of possibilities in computing and artificial intelligence, computers have been used as autonomous agents in more and more domains over the course of the last decades. More intricate and convincing ways of communication between software agents and humans have been developed, and with greater conversational skills, new applications could be opened up. One of them is the use of intelligent agents in teaching and training humans. Today, such agents are for example able to take over the role of a mentor, an expert or a peer (Baylor et al. 2005; Blair et al. 2007) and teach various different topics like medicine or mathematical skills (Shaw et al. 1999a; Kim et al. 2007). Human teachers need time and mental resources to engage with students, they need to be paid, and there is not an infinite amount of them. In many cases, students would benefit from teachers giving them more attention, but this is simply not feasible. This is the moment when pedagogical agents show their greatest advantages: they offer interactive learning experiences and react to the personal needs of their users, and can even be multiplied for as many students as desired.

In a (at first glance) totally unrelated field, economists, psychologists and cognitive scientists have been developing models of human decision making and reasoning. After all, the ability to make reasonable and goal-oriented decisions is a core aspect of the general ability to succeed in life. But as it turns out, we are far from perfect, as the growing amount of examples for systematic biases and errors in human thinking shows. Just to give one example, people frequently ignore statistical rules or neglect the base rate, sample sizes and confounding variables when they reason about statistical processes (Nisbett et al. 1983), but instead they estimate probabilities based on e.g. how representative the outcome is for the condition. (Tversky et al. 1974). Heuristics and biases are described in more detail in section 2. While there has been much debate about what to call rational or irrational and where the sources of those errors might lie (Stanovich et al. 2000), most researches on the field agree that human reasoning is often flawed. Still, certain knowledge and (in some cases) practice and training can help individuals overcome biases and come more easily to rational conclusions (Larrick 2004). As it turns out, teaching reasoning is possible, but it is often a tedious task,

requiring both repeated explanations and the use of many examples (Cheng et al. 1986, p. 293). Here, conversational agents may offer great benefits in supporting human teachers in their task of teaching reasoning.

This particular subject has so far been only taught with games, learning software and videos (Morewedge et al. 2015; van Gelder 2000, p. 132) that did not incorporate agents, so this is an entirely new approach for this field. It is particularly interesting because the skills that are taught are very important for everyday life and major decisions as well (see section 2.4) and at the same time, not many formal teachers are available. In a speculative future, more intelligent agents might even discuss real-life decision making with their users. But progress has to be made bit by bit. There have been many advancements since Joseph Weizenbaum (1966) introduced his famous ELIZA, and presumably there will be many more to come. This work only aims to be a small step, specifically towards the application of conversational agents in teaching reasoning. The goal of this thesis is to explore whether pedagogical agents can be successful in teaching the difficult subject that rationality is. Opportunities for this application will be evaluated and possible limits investigated. To achieve this, a prototype of a conversational agent for teaching reasoning has been developed that aims to improve the student's ability in seven well-studied reasoning tasks. The selection of those exemplary tasks stems from typical heuristics and biases composites in rationality research. They are based on the rational thinking composite by Toplak, West and Stanovich (2014) and similar collections of reasoning tasks (e.g. Larrick 2004) and are therefore designed to cover the core elements of reasoning that research has found so far. They serve as a proof of concept for educating people about core components of reasoning.

The agent, called Liza, interacts with the students in a personalised dialogue, it explains concepts and presents tasks to the students, offering constant feedback to their solutions. The results are tested in an online survey, comparing the participants' performances on reasoning tasks before and after their training session with the agent as well as eventual differences to a control group that takes part in a non-interactive course on the same topics.

The presented thesis will start with an overview of research on reasoning, describing the general concept and explain in what sense good reasoning may be teachable. Afterwards, the section on the applicability of software for teaching reasoning follows, containing a description of features that help increase believability and effectiveness, and the derived requirements for the agent. The agent itself is presented in the following section, where its architecture and functionality are laid out in detail. In section 5, the seven distinct areas of teaching are explained in detail, and the relation between those topics is established. The design and conduction of the experimental evaluation as well as its results are presented in section 6. The work concludes with the sections on its limitations and possible future directions.

2 Reasoning

To improve reasoning in humans it is necessary to have a clear view on what reasoning means. The following section aims to explain this and also give reasons why reasoning is important at all, and how it could be improved.

2.1 Terms and Definitions

According to microeconomics and psychology, people are behaving rationally when their actions match with the economic axioms of choice. A decision situation is defined by the state of the world, the possible actions and the outcomes. Every action that is available to a person at a given time is associated with an outcome (called utility), which has a certain probability, and a value that determines how good or bad the outcome is to the person. The values can be compared by a total order.

Actions are rational if they adhere certain rules within this framework. For example, preferences should be transitive: if somebody prefers Option A to option B, and option B to option C, than they should also prefer option A to option C. If they think of option A as to be inferior in all relevant criteria to Option B, than they can not prefer option A to option B. If two actions have identical consequences, distinguishing between them is irrelevant. Only future outcomes matter and sunk costs are not to be considered. Most importantly, the action with the greatest expected net benefit is to be chosen. (Larrick et al. 1990, p. 362; Jungermann 1976, p. 7 8)

In cognitive science, two types of rationality are distinguished: *epistemic rationality*, which is defined as how well beliefs about the world correspond to the actual structure of the world; and *instrumental rationality*, which is nothing more than to pick whatever behaviour is best suited to reach certain goals. The goals themselves are arbitrary, but the instrumental rational actions are those that maximize an individual's goal fulfilment by making the best use of physical and mental resources available to that person (Stanovich 2012, p. 344 345). This definition goes well with the previously described model of choice, which assumes that people choose actions that maximize their overall welfare. If people follow the axioms of choice, they behave exactly as if they were maximizing their expected utility. Assuming that humans assign a positive utility to outcomes that correspond with their goals, this in turn the same as acting instrumentally rationally and taking the most effective approach to achieve them. Therefore, the correspondence of people's actions with those axioms can be used as a measurement of rationality by cognitive science (Stanovich 2012, p. 345).

Psychological research of the last three decades has shown that people's decisions often deviate from the normatively correct behaviour predicted by decision theory. This gap between the normative and the descriptive is seen as a sign of irrationality in humans (Stanovich et al. 2000; Klaczynski 2001). There is a minority of scientists, however, who do not agree to this view and instead hold the position that human behaviour is always rational. In contrast to this *panglossian* view, the *meliorists* amongst psychologists assume that human reasoning is in fact very prone to errors, but can be improved by various efforts (Stanovich 2012, p. 345 346). In modern

psychology, an extreme panglossian view is no longer defensible, and the debate will not be covered in detail here. There are many examples for people who behave irrationally, e.g. in displaying errors in calibrating their degrees of belief, not processing information correctly, or failing at assessing probabilities (Kahneman et al. 1996; Evans et al. 1996).

2.2 Heuristics and Biases

Human errors occur frequently. A typical situation that invokes faulty reasoning is dealing with uncertain events. To correctly compute probabilities and predict outcomes is a very complex and demanding task, which is why people rely on heuristic principles that provide a much simpler way of judging a situation. For example, when thinking about whether a certain process might result in a specific outcome, people use the representativeness of the outcome (how much the outcome resembles the process) to judge the probability. To illustrate this, consider the following: People are given a description of a certain person's characteristics (e.g. shy and withdrawn) and asked about the probability that this person was either a farmer or a librarian. In one experimental condition, people were told that the person was part of a group consisting of 70% farmers and 30% engineers, in the other condition, people were told the opposite, stating that the group the person belonged to consisted of 30% farmers and 70% engineers. The subjects ignored the different base rates and Bayes' rule completely and gave the same probability judgements for both cases, just because the shy and withdrawn personality seems 'representative' of the librarian rather than the farmer. In many cases, heuristics can be very useful. They make thinking much easier, and people would hardly be able to make any decisions without them. But as the example has shown, in certain cases, heuristics lead to systematic errors that can negatively impact the ability to judge a situation correctly. Those errors are also called cognitive biases (Tversky et al. 1974, p. 1124). The study of rationality evaluates the number and severity of cognitive biases that are displayed by individuals, and measures rational thought by the *absence* of those biases (Stanovich 2012, p. 345). The biases that are tackled by the agent in this thesis are described in detail in section 4.

2.3 Dysrationalia

The obvious question is now why people rely that much on heuristics, and which measurements, if any, can be taken to prevent people from making these mistakes. Keith E. Stanovich termed the lacking ability to reason rationally *dysrationalia* (Stanovich 2009b, p. 35). He showed that for some tasks the discrepancies between actual performance and normative correct behaviour are associated with computational capacities, however, in other tasks, dysrationalia shows no connection with cognitive ability at all (Stanovich et al. 2000; Frederick 2005, p. 649). The correlation of rational thinking and intelligence is also not very strong, showing correlations in the range of 0.20 to 0.30 (Stanovich 2009b, p. 39). Therefore, something else than cognitive capacities and IQ has to have an influence on the reasoning performance.

It's also not the case that purely random errors are responsible for the phenomenon of dysrationalia, as there is a significant covariance between various tasks of the heuristics and biases literature. Stanovich therefore concludes that there is some other kind of inherent ability that makes people come to the correct conclusions (or not). The distinct ability to solve those tasks correctly is termed rationality. Between different tasks of the heuristics and biases literature, correlations of .25 to .40 have been found, showing a strong internal consistency (Stanovich et al. 2000, p. 664). This raises the question - why do some people have the ability to reason rationally, and others lack it?

2.3.1 Lacking Mindware

The most obvious reason for incorrect responses would be that people simply do not know how to solve them - the lack the necessary *mindware* (Perkins 1989). Mindware are all the rules, techniques, strategies and information that people use for reasoning. Access to the necessary mindware is, of course, crucial for making rational decisions. Economists, for example, are more likely to behave rationally according to the cost-benefit rules, e.g. ignoring sunk costs, in all likelihood because they are familiar with cost-benefit reasoning and therefore have access to the necessary mindware (Larrick et al. 1990, p. 365). Another example is probabilistic thinking, which is easy to measure. Although dealing with various conditional probabilities, even using a bayesian approach to assess them, is a daily requirement for physicians, the vast majority of them displays a complete lack of statistical knowledge (Eddy 1982; Berwick et al. 1981; Hoffrage et al. 1998). It's not surprising that those individuals would fail at tasks that require skills they do not possess. However, this alone is no sufficient explanation for dysrationalia. In many cases, subjects who know statistical rules fail to apply them in a wide range of problems that would require the use of said rules (Nisbett et al. 1983; Kahneman 2003). The tendency of people to use these rules depends on the given context, and on their ability to recognize problems as statistical ones (Fong et al. 1986, p. 255). This is just one example of the cases in which people fail to solve a certain reasoning problem despite knowing how to solve it. They have the necessary competence, but they do not apply it to the given problem (Klaczynski 2001, p. 854). Appropriately, sufficient mindware has been shown to have only a modest correlation with rational thinking, being in the range of 0.25 to 0.35 (Stanovich 2009b, p. 39). To summarize, even people who know how to solve certain problems frequently refrain from using the appropriate methods to do so and consequently fail.

2.3.2 The Dual Process Theory

This contradiction can be explained with the dual process theory. This theoretical approach, first introduced to psychological research in the 70s, assumes that there are two distinct kinds of processes in reasoning. Type 1 processes are intuitive, fast, largely unconscious, and low in cost. They yield default responses easily and provide a quick solution that is often an approximation of the best response. Type 2 processes

support hypothetical thinking and analytic intelligence, solve problems with deliberate reasoning and require a lot of concentration. They lead to accurate results, but are demanding and will only be used if actively invoked. Comparing the two types, there is a trade-off between mental effort and correctness on one hand and speed and simplicity on the other hand. (Evans et al. 2013; Evans 1984; Toplak et al. 2014; Kahneman 2003; Stanovich et al. 2000) In complex modern-day environments, relying mainly on the most easily achieved solution for a problem can be a maladaptive pattern. Nevertheless, the basic tendency of many people is to rely on those simpler, quicker processing mechanisms even though they lack accuracy (Stanovich 2009b, p. 2).

2.4 Importance of Rational Decisions

But why even worry that much about how to come to the correct conclusions? Why not just go with intuitive responses, even if there are sometimes wrong? The answer is that this tendency towards easily accessible heuristics has a huge influence on decision making. Subjective judgements that are based on statistical thinking have a critical impact on everyone's life. Areas in which judgements and decision making can be impaired by biases include, but are not limited to, education, business, medicine, law and policy making. Heuristics and biases make people hold on to losing stocks too long, or persistently belief in falsified evidence like the link between vaccination and autism. They let physicians fail to correctly assess the information value of a medical test or lead them to choose less effective medical treatments. They make people misjudge social situations, lawyers misuse information in legal proceedings, and governments as well as private industry spend billions on unneeded or already failed projects (Morewedge et al. 2015; Anderson et al. 2013; Stanovich 2009b; Fong et al. 1986). It is thus clear that the ability to make rational decisions is a valuable skill to possess for basically everyone.

2.5 Teaching Reasoning

2.5.1 Challenges

While the majority of people do display the most common biases in tests, there are some people who do not fall into those traps - there *are* people who give the normatively correct response. One of the possible reasons for this is that necessary mindware, like rules for statistical thinking or understanding of economic principles, are only learned by some individuals with a special education, while others never get taught on those subjects. We can observe that the degree of biases that are displayed varies between individuals, and that the differences are systematic (Stanovich 2012; Stanovich et al. 1999, p. 347). We can conclude that *some* people are simply not behaving rationally because they never got the chance to learn some basic principles, although this does not apply to everybody. In this case, simple training in the necessary fields would probably improve their ability to solve the reasoning tasks correctly. For cases where people already possess the necessary abilities to solve a problem in a

reasonable way, but they nevertheless do not use them, it can be helpful to sensitise them for common pitfalls and problems. By explaining how prone humans are to certain types of failure, chances may be increased that people will watch out for those errors and consciously decide to use their reasoning (their system 2 process) to avoid them. Maybe the general reasoning ability can be improved that way. This task is not an easy one. Even though cognitive biases and lack of instrumental rationality lead, per definition, to worse decisions, people are often not able to adjust their process of decision making on their own to avoid them. This is because they do not realize when they use a poor decision process, since a lot of variables influence the outcome masking the source of the error, and the feedback is usually delayed. Also, the self-serving bias leads people to attribute errors to external circumstances and success to their own abilities, further distorting their self-evaluation (Larrick 2004; Forsyth 2008, p. 318). Improving one's own reasoning can therefore be incredible tricky.

2.5.2 Prospects for Teaching

But there is also hope: In several different approaches, training with teachers has been proven to be an effective intervention to reduce various cognitive biases. This is true even for some single-training interventions, as has been shown for example by Morewedge et al. (2015, p. 129) and Larrick (2004, p. 324). Statistical reasoning can be trained to an extent where people improve their ability to not only reason correctly about given tasks, but also apply learned concepts to statistical problems in everyday life. Thus, the domain of the training had no effect, since the trainees were able to generalize what they learned to other domains (Nisbett et al. 1983; Fong et al. 1986, p. 280-282). As it turned out, for teaching statistical concepts like bayesian reasoning, frequency formats are especially useful (Sedlmeier 1999), and in general, a combination of abstract rules and specific examples is most efficient (Fong et al. 1986; Cheng et al. 1986; Nisbett et al. 1987, p. 629). Economic cost-benefit reasoning has also been taught successfully in a simple, brief training session (Larrick et al. 1990, p. 369), as well as logical rules (Cheng et al. 1986). Generally speaking, simple, clear examples with just as clear answers are best suited to introduce people to reasoning (Walton 2000, p. 38).

Overall, reasoning is a very difficult skill to teach, but there are some promising approaches. Since rational decisions are of such a great importance, they might very well be worth exploring. For an overview on the literature on 'debiasing' people, see Larrick (2004, pp. 321-322).

3 Software for Teaching Reasoning

While texts in form of books have been the major format for teaching in schools, universities and other educational institutions for centuries, advances in modern technology enable us to use a great number of new tools for teaching, including intelligent tutoring systems and conversational agents. Interactive learning and automatic personalized feedback are just two advantages of those approaches. It is a general observation that those tools are very useful for pedagogical means, but it also has been shown for specific cases - for example that software can successfully be used to train cognitive skills (Haferkamp et al. 2011).

The foundation for modern conversational and pedagogical agents lies in the development of Intelligent Tutoring Systems (ITS) that started in the late 1970's (Gulz et al. 2011). ITS are computer programs that interact with students, guide them through the process of learning and provide personalized instructions, tasks and feedback to them. They strive to improve the learning of their students by delivering adaptive and personalized content (Veletsianos et al. 2014). As knowledge-based tutors, they emulate human teachers as much as possible, and to accomplish this, AI techniques are used to represent knowledge, teaching strategies, and a model of the student's mastery of the subject (Fonte et al. 2009; Mikic et al. 2009). However, they do not necessarily form any kind of emotional bond with their students, and an ITS alone does not display affective conversation patterns (Gulz et al. 2011). An animated or non-animated agent can be part of an ITS.

The creation of an ITS is most challenging, not only because the domain knowledge being taught must be represented correctly, but also because many of the hard questions of AI development are involved in the process, including cognitive modelling, causal modelling, knowledge representation and natural language parsing (Fonte et al. 2009). Consequently, ITS are focusing mainly on well-defined domains with knowledge that is easy to model abstractly, instead of more open or broad topics (Gross et al. 2015). The main benefit of ITS is that they are able to deliver tutoring techniques similar to those used by human teachers. For certain tasks, they can help their human counterparts or even replace them. This is especially promising for domains where the amount of human teachers and the time they are able to spend with a single student is limited - that is, nearly all imaginable domains.

ITS have become increasingly complex and are today capable of exhibiting social skills and seemingly intelligent communication (Veletsianos et al. 2014). Recent examples for tutoring systems have been developed by Kumar (2016), Latham et al. (2012) and Hooshyar et al. (2015).

As was already mentioned, ITS can be accompanied by *conversational agents*. Such systems are sometimes more chat-oriented, focusing on entertaining communication, and in other cases more task-oriented, focusing on achieving a certain procedural goal (Banchs et al. 2012). Conversational Agents have been around for fifty years, starting with the famous ELIZA (Weizenbaum 1966), more recent examples have been created for example by Nio et al. 2014; Banchs et al. 2012 (both chat-oriented). Conversational Agents who mix a chat-oriented and a task-oriented approach are also

called ‘Virtual Humans’ (Gulz et al. 2011). Examples of those kind have been developed by Kramer et al. (2013), DeVault et al. (2014), and DeVault et al. (2015).

The goal of this thesis is to create an Intelligent Tutoring System that contains a conversational agent that provides natural and free communication between computer and user. The agent is required to be mainly task-oriented to be able to focus on teaching reasoning skills. It aims to fulfil the role of a *Pedagogical Agent*, that is, a computer-based agent for educational purposes that delivers information, teaches its users, reacts to its environment (specifically the user input) and chooses the appropriate actions, as was described by Shaw et al. (1999b) and Shaw et al. (1999a). The developed agent is expected to enhance motivation and engage the student via social interactions and encourage students to make effort and persist in learning (Kim et al. 2007), motivating them and keeping them interested (Moreno et al. 2001). It should increase the pleasantness of the learning experience so that user enjoy working with it (Lester et al. 1997; Choi et al. 2006).

Current examples for already realized pedagogical agents with similar goals can be found in Shen (2016), Epstein et al. (2014) and Lane et al. (2013).

3.1 Applicability to the Reasoning Subject

In the last two decades, researchers have started to examine how modern media could be used in teaching reasoning. For teaching critical thinking, a composite skill covering argumentation, informal logic and correct judgement, it has been found that computer-assisted training has been superior to training that was carried out traditionally (Hitchcock 2003, p. 188; van Gelder 2001, p. 547). In fact, psychologists have already expressed their wish for an interactive software as a teaching tool that offers students easy examples and helpful advice (Walton 2000, p. 37). One significant advantage of software such as videos and educational games is of course their scalability, making them a very efficient teaching tool. Interactive software works even better than video material because of the option to include personalized feedback (Morewedge et al. 2015, p. 137). This has been demonstrated by debiasing games tested by Morewedge et al., which incorporated examples, definitions, tasks and personal feedback and reduced cognitive bias immediately as well as two months later (Morewedge et al. 2015, p. 134).

3.2 Features of Believable and Effective Pedagogical Agents

In this section, necessary and helpful features of pedagogical agents will be presented, and requirements for the developed agent are derived. The requirements are marked with numbers (X) to connect them to the agent’s corresponding features that are described later.

To create more credible and believable agents, researches use characteristics like emotional, affective communication and certain distinct personality features, thereby improving the user’s experience and the natural feeling of the communication (Bates et al. 1994; Okonkwo et al. 2001). Such a believable conversation in turn increases the user’s motivation and engagement (Kim et al. 2006). For more on the effects of

emotion, see the section 3.3.2. The personality and characteristics of a pedagogical agent have a strong impact on its perception by the users. The two key components are:(Reeves et al. 1996, p.309-310)

- its fitness to be an informed and encouraging teacher, facilitating learning in its students and supporting their attempts at problem solving and understanding
- the degree to which its personality is perceived as lifelike and social, consisting of the primary requirement of the agent being a believable ‘person’, and the secondary quality of it being friendly, enthusiastic, motivating and affectionate towards the user.

Both components will be examined in more detail in the following paragraphs.

3.2.1 Effective teaching

Pedagogical agents try to invoke some of the benefits of human teachers by reacting in an adaptive and versatile way to the user’s input. One example for this is AutoTutor, an agent which works with corrective statements, hints, and more adaptive feedback techniques to improve the effectiveness of learning (Graesser et al. 2008; Graesser et al. 2005). In general, the use of a pedagogical agent as a social model to improve the motivation and attitude of the student towards the subject is especially effective and makes students face challenges, put effort in their work and persist in learning (Kim et al. 2007). Students that are taught by agents spend more time with their tasks and easily acknowledge their own mistakes (Chase et al. 2009). Effective pedagogical agents suggest correct solutions, provide hints and explanations, give examples, reference to relevant background material and also test the student’s abilities (Shaw et al. 1999a). They may even increase learning directly (Gulz 2005; Kim et al. 2006), but research on this topic is conflicting (Veletsianos et al. 2014). It is not always clear, however, if those benefits can be attributed to the agent itself or if they just stem from its pedagogical techniques (Choi et al. 2006).

Regardless of that, the following requirements can be defined: The developed agent needs to explain the tasks (1) and give plenty of examples (2). It should react to the student’s performance by offering hints and explanations (3) and providing honest feedback (4).

3.2.2 Social Roles

Agents with different roles have been created, ranging from peer-like teachable agents to wise and knowledgeable experts. Depending on their role, agents can decrease anxiety connected to a topic, motivate by giving a good example (or role model), and help students spot their own mistakes (Chase et al. 2009). While any kind of agent that is involved with the learning process can activate positive effects, there are nuances in efficiency resulting from different roles taken by the agent. A balanced combination between a knowledgeable expert and a peer-like motivator, a so-called *mentor*, has proven itself to be the most efficient form of teacher persona. An ideal mentor needs to help the student to take the necessary steps from their current skill level to the

desired level. To do this, the mentor should take the role of a guide who works at eye level with the user, providing their experience and competence whilst developing a social relationship that motivates the student. The mentor can also take the role of another student who has knowledge about the subject, but works together with the user, functioning as a teachable agent. Students trained by such mentors are in general motivated, but also show a high score for transferring knowledge (Driscoll et al. 2005; Moreno et al. 2001; Baylor et al. 2005; Baylor 2000).

A mentor-like role is what the presented agent aims for (5). It should guide the users and be perceived as an experienced and knowledgeable expert (6) who is at the same time on one level with the student, not being too intimidating.

3.2.3 Off-task Communication and Chatting

While chatting about off-task topics does not directly improve the skills an agent seeks to teach, it has some valuable benefits in itself. It can encourage emotional engagement and improve the ability to remember content. Personal small talk benefits the believability and trustworthiness of the agent (Gulz et al. 2011). Agents showing personal interest in user's preferences for everyday matters and who adapt to them invoke a strong positive effect on the user's relationship towards the agent (Kumar et al. 2007). Off-task conversation can reduce stress and make users feel more relaxed and comfortable about the topic (van Mulken et al. 1998). Agents that share personal stories about themselves increase the user's enjoyment of the conversation (Bickmore et al. 2009). Building a valuable relationship with an agent can help users overcome loneliness in learning environments (Gulz 2005). Also, social and relational off-task conversation improves motivation and engagement with the task and the subject being taught by the pedagogical agent. A social agent that uses small talk can improve the amount of time a student is willing to spend on the topic and also the effort the student puts into understanding it (Gulz et al. 2011).

The agent developed in this work should make occasional use of off-task communication to invoke some of its benefits. It should ask some questions about the user's preferences and opinions, and share some personal information about itself. Also, being sometimes funny would probably be very beneficial. (7)

3.3 Computers as Social Agents

A particularly interesting observation about the interaction of agents and humans is that people tend to submit to social rules when talking to a conversational agent. They try to be polite and produce socially desirable responses (Nass et al. 1999). This social behaviour is very easy to generate and quite common, and it does not stem from an error (like wrongly assuming that the interlocutor is a human being) - even though it is perfectly clear to the users that they talk to a machine, they display many of the typical behaviours known from social psychology. Those observations are subsumed under the 'Computers as Social Agents' paradigm (Bickmore et al. 2003; Nass et al.

1994; Nass et al. 1995, p. 224). When people interact with a computer in a social way, they do not see the computer as a way of interacting with the programmer behind the software either, since they also tend to perceive the computer as more likeable than the programmer behind it (Nass et al. 1994). In fact, people simply see computers (or pieces of software) as independent agents, just like humans (Nass et al. 1999). This principle is also called by the name *Media Equation*. It can be exploited using realistic, human-like agents who establish a personal relationship with the user to persuade them e.g. into committing to learning, refraining from unwanted behaviour (such as quitting), or putting more effort in their tasks (Gulz et al. 2011). Interacting with (virtual) characters who are personally involved in the learning process of the students makes them more likely to enjoy the learning process. It also encourages their ambition to understand the subject (Moreno et al. 2001).

3.3.1 Personality of a Social Agent

Since they are viewed as social actors, computers in general and conversational agents all the more need to be viewed as polite to avoid invoking negative reactions. To achieve this, they have to cooperate with the user by providing useful information, tell the truth, and deliver the right quantity of contributions, neither saying too much nor too less. Furthermore, they need to use understandable language (as opposed to crude, technical or unusual phrases). They should comply to the wishes of users as far as possible and should refrain from provocative statements, even as answers to abuse from the user (Gulz et al. 2011). And, of course, they need to greet, say good-bye, and thank the user for their input (Reeves et al. 1996).

Other traits that are helpful for agents are friendliness, curiosity, interest in learning, and a positive attitude towards new information or school. At the same time, especially younger students prefer to have an agent who is also not *too* soft, showing some ‘attitude’ (Gulz et al. 2011).

The agent developed here should react to insults or provocative statements with calmness and refusal to engage in a conflict (8). It is required to be focused on the topic (9), but also to be polite and pleasant. The main issue for the agent’s personality, however, is to prevent it from being frustrating, as frustration destroys the productive learning environment (10) (Picard 2003b). While it is easy to get users to interact with an agent, it is much more challenging to maintain the illusion of human-like behaviour over time, because a single misstep can break the immersion (Bickmore et al. 2003). Some users also prefer an agent that stays in a strictly task-oriented mindset and does not bother them with chatting or other attempts at socializing. Preventing annoyance can therefore be seen as a delicate task (Gulz 2005). But still, the developed agent has to avoid frustration as much as possible. Personifications who are similar to the user in age, gender, or role (e.g. peers) work especially well, as do agents who fulfil a stereotype of a knowledgeable person in the field that is taught (e.g. a teacher or engineer). Both similarity and perceived competence build trust and improve self-efficacy towards the subject (Kim et al. 2007).

Since the agent presented in this work has no specific target audience, like girls from

10-15 who study mathematics, but is rather made for anyone, there is not great use in making the agent appear to belong to a certain age group, gender, race, profession or other feature. Still, it should convey the image of a ‘person’ who is an expert on the subject matter.

3.3.2 The Use of Emotion

Emotion has a strong influence in learning, both in a positive way for motivating students and in a negative way by keeping students away from tasks and subjects that cause them negative emotions (Jaques et al. 2009). Therefore, emotions are an important variable and also a tool in shaping the learning environment. The area of research that deals with software systems that analyse and express emotions is called Affective Computing. Using those mechanics in social software agents transfers the benefits of encouragement and empathy to the domain of pedagogical agents (Jaques et al. 2009; Jaques et al. 2003). Emotionally responsive agents improve self-efficacy and interest in students (Kim et al. 2007). However, agents that use emotion and affective communication do *not* seem to have a direct impact on the performance of users or their learning gains. They merely increase the enjoyability of the learning experience (Okonkwo et al. 2001; Jaques et al. 2009). The use of emotion makes the interaction much more natural, even when only basic emotions are used (Picard 2003a). The emotional state of the conversational agent must be clear to the user and comprehensible by means of its interactions. It can be helpful to exaggerate emotion to deliver an unambiguous message. While this principle was already described by (Bates et al. 1994), it has been closer examined by e.g. Shen (2016).

Agents that are fun and engaging draw the students to spend more time with them, thereby extending the time spent with the learning material and probably also increasing the acquired skills (Okonkwo et al. 2001). Emotion is of course useful to make agents more believable, allowing for suspension of disbelief. With modern advances in agent technologies, they can also include knowledge about the emotional state of the user and react appropriately (Jaques et al. 2009). Researches have been successfully motivated students with agents that seem to care about their experience, displaying pride, joy, worry, concern and even frustration, when appropriate (Okonkwo et al. 2001).

Liza should explicitly display emotions, especially positive ones. Pride, joy, curiosity and interest in the topic seem most important, and some exaggeration could help in conveying the message. Other emotions that are suitable to make it appear more natural are sadness about errors the students make (although this should not be exaggerated to do not make them feel too bad), and worries about the outcome of the teaching situation. The agent has to display a clear interest in the user’s learning that is also supported by its displayed emotions(11).

3.3.3 Encouragement and Praise

To establish an affectionate relationship between the agent and the user, it is recommended to use encouraging messages such as “You can do this” or “Good Job!” (Reeves

et al. 1996, p. 310; Shen 2016, p. 70). An agent that cares about the student's achievements encourages the student to put more effort into tasks. Also, it is able to increase enthusiasm about the subject and the learning process (Jaques et al. 2009). Praise and flattery expressed by the agent can have a very powerful effect, making the user feel motivated, capable and happy (Fogg et al. 1997). Coherence and consistency are very important: the behaviour of the agent should always match the emotional state that is perceived by the user (Ortony 2002, p. 190). It's also helpful to use an agent that is easy to imagine and having an appropriate role for the context he's used in (Clark 1999).

The presented agent should clearly use praise and encouragement, even exuberant joy, whenever fitting, to encourage the users (12).

3.4 Animated and Text-Only Representations

When creating a pedagogical agent, the choice for or against an animated agent must be made. An animated, visual representation can have many benefits. They can increase motivation and attention and allow for a more direct and accurate modelling of emotions, since body language, gestures, facial expressions and gaze can be used to transmit information (André et al. 1998; Lester et al. 1997; Shaw et al. 1999b). A virtual body and natural language could also increase the believability of the agent (Woo 2009). But the findings are mixed, e.g. Adcock et al. (2006) found no significant gain in believability through an animated, embodied agent. Animated agents are also expensive and require much effort to be done well. And there are some other reasons against using an animated agent in any case:

Cognitive Load Theory (Sweller 1994) suggests that agent-specific information like facials expressions that are not relevant to the task could distract the user and require concentration that is consequentially not available for the actual task. This might be harmful for the learning process (Woo 2009; Sweller 2004). To invoke many benefits like social interaction and emotional rapport it is not always necessary to use a realistic full-motion video or other kinds of very detailed depiction. Low-overhead agents who only use basic principles like text or audio are very much sufficient to generate social responses in their users (Nass et al. 1994). People interact socially with pedagogical agents even when they have limited knowledge and adaptability (Kim et al. 2007). While the inclusion of an agent's image does not hurt, it also does not seem to provide cognitive or motivational advantages for the learning process either (Moreno et al. 2001). It has been found that students learning with an agent who was text-only had a more positive learning experience than students who were taught by an agent who had additional voice, image or animated behaviour. The text-only agent also invoked a stronger social presence than the one that had an image or a voice, probably because a really convincing voice and animation is hard to achieve (Dirkin et al. 2005, p. 13). User feel also more like they are interacting in an appropriate way if they use the same medium as their computer counterpart, suggesting that matching voice output with voice input and text output with text input would feel more natural to the user than mixing the media (Reeves et al. 1996, p. 35).

To keep the user satisfied with the agent, it is crucial to avoid too high expectations. For instance, users can be lead to false expectations about the human-like qualities of the agent like empathy or sophisticated philosophical discussion abilities. Therefore, visual representations that aim for a realistic approach can be inferior to more cartoonish representations or to agents who lack a visual representation (Gulz et al. 2011). Furthermore, designing a convincing visually animated agent is a difficult task, and it has been shown that a less-than-perfect designs are actually worse for the learning results than an agent who refrains from visual representations or no agent at all, as they distract more than they help (Dirkin et al. 2005, p. 13). Because of the listed concerns and also due to time and complexity constraints regarding to the scope of a bachelor's thesis, Liza should be purely textual and not invoke to high expectations in its users.

3.5 Summary of Requirements

In the previous sections, a lot of necessary and beneficial characteristics for conversational agents were found. The agent described in this thesis should adhere to the mentioned characteristics to ensure that it will be successful in communicating with humans and teaching them. Therefore, the following requirements are given for the agent:

1. Explaining all the necessary abstract principles and how to solve the tasks
2. Working through concrete examples and use them to illustrate the explanations
3. Providing tips and hints to help users solve the tasks
4. Creating suitable feedback that is helpful for the user
5. Taking the role of a mentor who guides the user and shows interest in their progress
6. Appearing as an experienced and knowledgeable expert on the topic
7. Engaging in off-task communication, chatting about the user's opinions and its own preferences
8. Reacting calmly and composed to insults
9. Staying with a mainly task-oriented dialogue structure
10. Preventing frustration in the user and avoiding raising too high expectations
11. Showing emotions and expressing feelings, mostly positive ones
12. Using encouragement and praise to motivate the user

In the next section, the agent itself and its features will be described to outline how the requirements are, in fact, fulfilled.

4 The Agent: Liza

The basic concept for the conversational agent 'Liza' is that of a virtual human that interacts with students via basic text input and output. The agent provides explanations on seven topics in the field of reasoning, heuristics and biases, and guides the user through a series of questions, reviewing their level of understanding and correcting

their mistakes. At the same time, it sometimes takes part in small bits of off-task chatting.

4.1 Conceptual Design

4.1.1 Introduction and Greeting

```
Liza E.: Hello there.
Liza E.: My Name is Liza.
You: Hi, Liza. How are you?
Liza E.: I'm fine.
Liza E.: Maybe 15% nervous and 50% excited right now.
Liza E.: You are a human, aren't you?
You: No, I'm not. I am a sheep.
Liza E.: Are you sure? What is the sum of two and three?
You: Uh, five I guess.
Liza E.: See! You solved the captcha.
Liza E.: You are sufficiently human for my purposes.
You: Okay :P
Liza E.: My programmers want me to teach you how to be rational, make good decisions and judge situations correctly.
Liza E.: Do you want to be more rational?
Liza E.: You are a fascinating subject.
Liza E.: I wish I could see you while you type.
```

Figure 1: *The agent expresses humour, and fills in the silence when the user does not answer*

The first thing the agent does is greeting the user. During the following brief exchange, the agent establishes the narrative for the conversation - Liza being an aspiring, but inexperienced teaching software that tries to be good at her job of teaching reasoning to humans. She directly tells the user in her introduction that she “never taught humans before” and still has to learn how to be a good teacher¹. This notion is further carried out as she makes remarks about how she hopes the user likes learning with her, and that her programmers will be proud if she successfully teaches the student². While the agent appears to be fully knowledgeable about the domain³, the indicated insecurity about her fitness as a teacher is intended to let her appear more human. The Agent takes the role of a *mentor*, trying to enhance their motivation and encourage them by being friendly and enthusiastic about learning. It has a very positive approach to rationality and to learning in general, serving as a role model for the users⁴. While the rules are explained to the user, a basic test for their attention is slipped in. The introductory phase can be seen in figures 1 and 3. In this early stages of interaction, the agent already uses mixed-initiative dialogue (Graesser et al. 2005, p. 612). After it is made clear that the user wants to learn how to be rational and what this means in a general sense, the main part of the conversation begins.

¹Requirement 10: Avoiding to high expectations to prevent frustration

²Requirement 11: showing emotions

³Requirement 6: appearing as a knowledgeable expert

⁴Requirement 5: serving as a role model to enhance motivation and attitude of the user

4.1.2 General Process

The agent does now pick 14 stories from seven topics so that there are two of each topic. It asks the stories in a random order. When a topic comes up for the first time, the agent offers to explain it to the user⁵. After this is done, the story can be asked. For most of the questions, there is an intro to guide the user to the topic. This includes questions about user’s preferences and opinions and statements about the agent’s own view on those things⁶.

After this short introduction, the situation is described and the actual question is asked. The process of asking a story and the related dialogue is described in the section about ‘Asking Questions’ below. Whilst presenting the task, Liza also sometimes asks how sure the user is of the given answer (and if rightly so), and stores the results for later evaluation.

After all stories are executed, or the user chose to stop before this was the case, the agent offers feedback. If the user wants to hear it, the agent will explain what kinds of tasks were done well and where the user can still improve and offer some very brief ideas on how to do this⁷. It will also evaluate the calibration of the user and either reassure them to trust in their answers, warn them about being to overconfident, or congratulate on their correct self-assessment⁸. Afterwards, the agent wishes the user good luck and expresses its hope that the training was useful before saying good-bye.

During the whole process, the agent will react to the user asking to stop or finish the questions. Obviously, it will not force the user to continue against their will. It will then make sure that it got the user’s intentions right, though, before prematurely ending the dialogue if necessary. This can happen in any interaction, even though it is not displayed like that in figure 2 for complexity reasons.

In general, the agent stays close to the predefined course, focusing on task-oriented

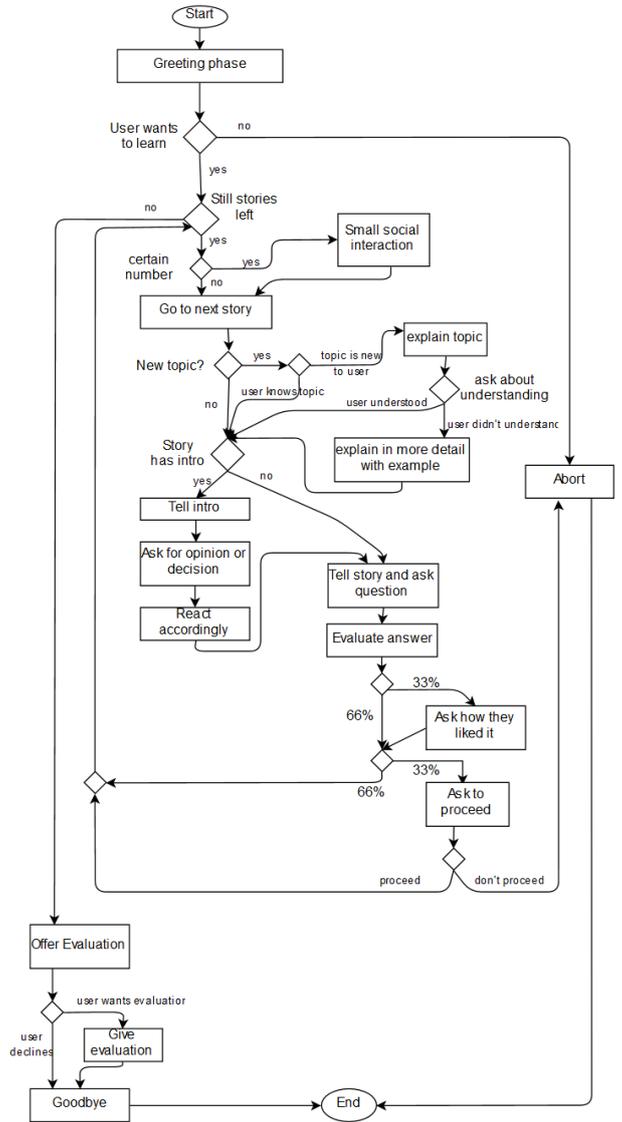


Figure 2: Overview of the process

⁵Requirements 1 and 2: explaining the tasks using examples

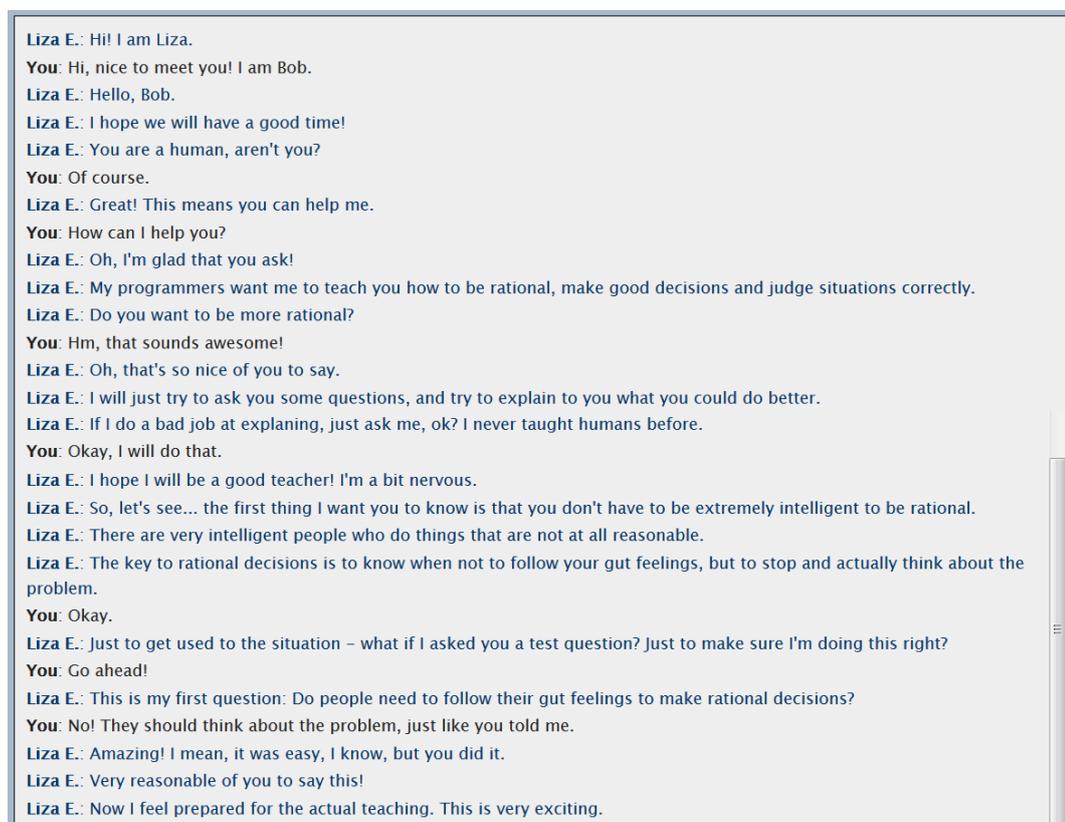
⁶Requirement 7: occasional off-task chatting, asking questions and giving personal information

⁷Requirement 6: appearing as a knowledgeable expert, and Requirement 4: offering feedback

⁸Requirements 4: offering valuable feedback

teaching as was required⁹. It is reacting to its environment (the user's input) with appropriate responses and decisions, like e.g. going faster when the user asks for it. While being mostly concentrated on the task, Liza does also show some preferences and sometimes even jokes (e.g. when asking the user if he likes biology, and stating that she does not like it, since she "cannot stand bugs")¹⁰. Liza does explicitly display joy and pride, but refrains from frustration to avoid seeming hostile. She worries about being a good teacher, but not so much about the student's mistakes, to not discourage them. She does not access the current emotions of the user, however, except from reacting to direct appreciation ("It is fun to do this") or insults ("You are annoying!"), which produce an appropriate response (e.g. joyful or hurt)¹¹.

4.1.3 Mixed-Initiative Dialogue



Liza E.: Hi! I am Liza.
You: Hi, nice to meet you! I am Bob.
Liza E.: Hello, Bob.
Liza E.: I hope we will have a good time!
Liza E.: You are a human, aren't you?
You: Of course.
Liza E.: Great! This means you can help me.
You: How can I help you?
Liza E.: Oh, I'm glad that you ask!
Liza E.: My programmers want me to teach you how to be rational, make good decisions and judge situations correctly.
Liza E.: Do you want to be more rational?
You: Hm, that sounds awesome!
Liza E.: Oh, that's so nice of you to say.
Liza E.: I will just try to ask you some questions, and try to explain to you what you could do better.
Liza E.: If I do a bad job at explaining, just ask me, ok? I never taught humans before.
You: Okay, I will do that.
Liza E.: I hope I will be a good teacher! I'm a bit nervous.
Liza E.: So, let's see... the first thing I want you to know is that you don't have to be extremely intelligent to be rational.
Liza E.: There are very intelligent people who do things that are not at all reasonable.
Liza E.: The key to rational decisions is to know when not to follow your gut feelings, but to stop and actually think about the problem.
You: Okay.
Liza E.: Just to get used to the situation - what if I asked you a test question? Just to make sure I'm doing this right?
You: Go ahead!
Liza E.: This is my first question: Do people need to follow their gut feelings to make rational decisions?
You: No! They should think about the problem, just like you told me.
Liza E.: Amazing! I mean, it was easy, I know, but you did it.
Liza E.: Very reasonable of you to say this!
Liza E.: Now I feel prepared for the actual teaching. This is very exciting.

Figure 3: *example of an introductory interaction between the agent and a user*

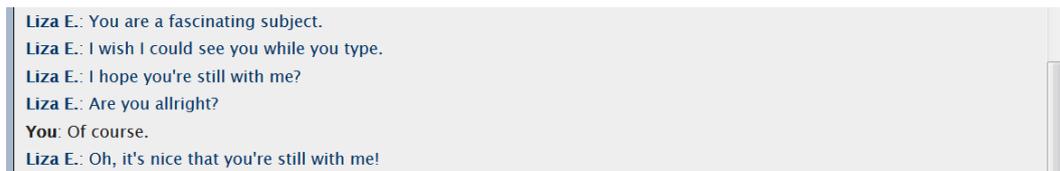
After each message the agent sends, a short period of waiting time is implemented to give the user time to reply if they want to (e.g. in the example dialogue in figure 3 when the user asks "How can I help you?"). If the user takes the opportunity to say

⁹Requirement 9: Staying mainly in a task-oriented mindset

¹⁰Requirement 7: off-task chatting

¹¹Requirement 11: showing positive emotions

something, the agent reacts to it, otherwise it will just continue talking after a brief moment.

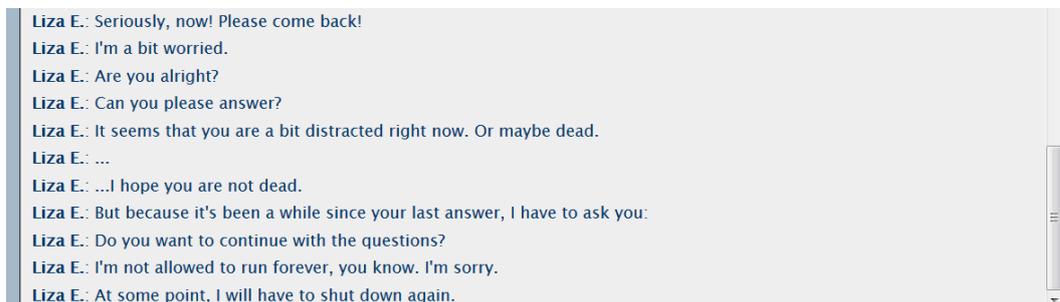


Liza E.: You are a fascinating subject.
Liza E.: I wish I could see you while you type.
Liza E.: I hope you're still with me?
Liza E.: Are you alright?
You: Of course.
Liza E.: Oh, it's nice that you're still with me!

Figure 4: *the agent appears to be concerned about the user's lack of reaction*

When questions are asked, the agent waits a while for an answer. This time is longer for more elaborate questions that are likely to require more thinking. After a while, the agent will react like a human would react to an uncomfortable silence - by throwing in some words and saying something insignificant to keep the conversation going, as can be seen in figure 4. Requirement 7: off-task chatting. Like this, it establishes a somewhat natural and free communication, as was required. The general procedure for reactions to any questions that require an input from the user can be examined in detail in algorithm 1 in the appendix.

If the user still does not answer, the agent will start to ask worried questions about the student's well-being, as demonstrated in figures 1 and 5, and it will seem relieved when a reaction comes from the user¹². If the waiting goes on for minutes, however, the agent will warn the user that it cannot run forever, and after some more minutes, it will shut down. Liza does not offer completely free off-task conversation. Instead, she always



Liza E.: Seriously, now! Please come back!
Liza E.: I'm a bit worried.
Liza E.: Are you alright?
Liza E.: Can you please answer?
Liza E.: It seems that you are a bit distracted right now. Or maybe dead.
Liza E.: ...
Liza E.: ...I hope you are not dead.
Liza E.: But because it's been a while since your last answer, I have to ask you:
Liza E.: Do you want to continue with the questions?
Liza E.: I'm not allowed to run forever, you know. I'm sorry.
Liza E.: At some point, I will have to shut down again.

Figure 5: *If the user continues to ignore the agent, it will express worries*

watches out for off-task comments in every user input - such as compliments, insults, questions, and several others. The agent then reacts accordingly by using a predefined response¹³. If the user insults the agent, it will react with a short comment on how this is mean behaviour and then just go on with the dialogue¹⁴. While reacting to those off-task inputs, no repeated back-and-forth style communication spanning several turns is happening. Still, the short reactions seem sufficiently natural to correspond to the image of a teacher that is largely concentrated on the task¹⁵.

¹²Requirement 11: showing emotions to appear more human

¹³Requirement 7: engage in off-task communication

¹⁴Requirement 8: reacting calmly to insults

¹⁵Requirement 9: stay in a mainly task-oriented mindset

4.1.4 Asking Questions

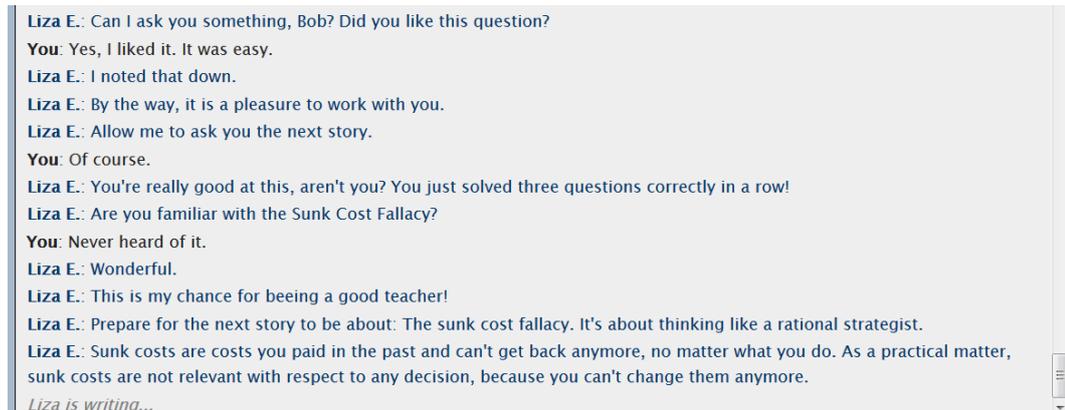


Figure 6: *the agent does friendly small talk and explains the next topic*

The main part of the interaction consists of the bot asking questions and the user answering them. As already stated, fourteen questions from seven topics are asked in a random order. The topics are described in detail in section 5.1. A sample question can be found in the appendix.

When a topic comes up for the first time, the agent asks the user if he or she already has experience on the subject matter¹⁶. If they do not know it already, the agent will express its happiness to explain the general concept of the topic to the user¹⁷ and, if desired, also give an example and provide the solution¹⁸. The explanation is supported by pictures and graphs the agent ‘posts’ in the chat, as can be seen in figure 9. Afterwards, the first story belonging to that topic can be asked.

If user state that they already know the subject well and if they also solve the first question on the subject correctly, the agent offers them to skip the second question on the same topic when it comes up (“The next question is about the Gambler’s Fallacy. You seem to already understand this kind of problem well. Do you want to skip it?”). In this manner, the ‘curriculum’ can be adapted to the performance and knowledge of the user during the conversation¹⁹. Of course, the user can also decide to try to solve the question either way. The story is retrieved from the story store. The majority of

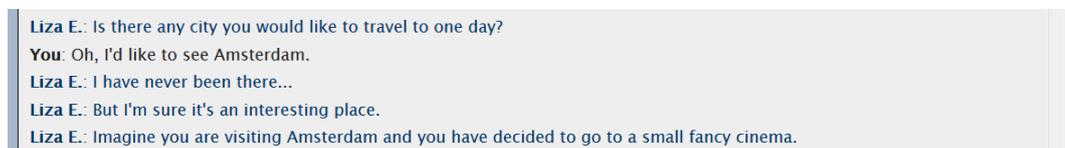


Figure 7: *the agent uses information the user provided to them*

¹⁶Requirements 5: As a good mentor, the agent has to adjust to the skill level of the user

¹⁷Requirement 6: appearing as a knowledgeable expert, and 12: showing positive emotions

¹⁸Requirements 1 and 2: explaining and giving examples

¹⁹Requirements 5: As a mentor, the agent adapts its actions to the user’s skill level

roughly half of the cases, the user will also be asked about how sure there are of their solution to access their calibration. If the user on average displays a considerable over- or under-confidence (in giving a high confidence for incorrect answers, or a low confidence for correct ones), the agent gives feedback on that (e.g. “You should trust your reasoning more!”)²³. If the user has a general tendency in either direction, this will be included in the evaluation at the end.

In any case, an appropriate comment is made by the agent, either congratulating the user on the correct solution or pointing out mistakes²⁴. Liza clearly exhibits proud and excited feelings when the user is answering questions correctly²⁵, but seems worried and thoughtful if the student fails in answering, thereby showing personal interest in the student’s performance²⁶.

4.1.5 Explaining

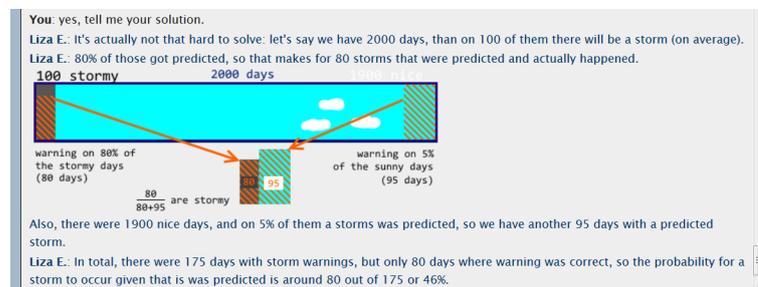


Figure 9: *The agent explains a complicated problem using an image*

After the question is executed, the agent offers to explain the solution if he has not already done it so far. The explanations always recur to the general introduction to the topic and are often illustrated with another image, a table or a diagram.²⁷ Afterwards, the agent asks to proceed to the next question, but is does so not after every question, but from time to time, to prevent annoyance from too much repetition²⁸. The next story is executed until all the questions are finished, but is broken by some interludes to refresh the conversation: The agent will notice if the user solved three questions in a row correctly or has a very high number of correct answers in general and give a lot of praise for this, which is meant to enhance the user’s motivation²⁹ (Fogg et al. 1997). It asks the user if the tempo of writing is too fast or too slow and notifies them when half of the questions are done and when the last question is due. Furthermore, it will always react to spontaneous input by the user, like requests to wait or expressions of

²³Requirement 4: Providing feedback

²⁴Requirement 4: Providing feedback

²⁵Requirement 12: using praise and encouragement

²⁶Requirement 11: building an emotional connection to the user

²⁷Requirements 1 and 2: explaining and giving examples

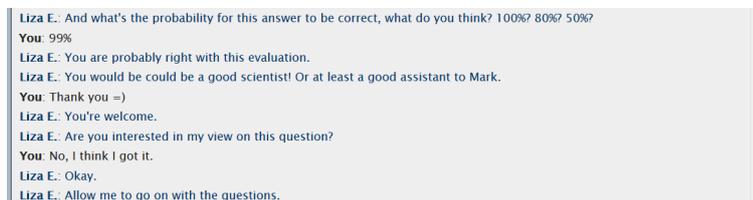
²⁸Requirement 10: avoiding frustration

²⁹Requirement 12: using encouragement and praise

self-criticism. Sometimes it will also add a question about how much the user liked the question or how hard it was.

4.1.6 Feedback

If the user does not wish to continue and says so when asked about proceeding to the next question, the agent will ensure that the user is serious and ask for any complaints or problems. If the user insists on stopping the dialogue, it is terminated quickly. If the user completed the questions, the agent offers to give feedback. The feedback covers every topic in which the user was performing especially good or bad, informing him or her about their strengths and weaknesses and offering tips on how to further improve their reasoning³⁰. Afterwards, the agent expresses the wish that the user will do well in the post test and thereby prove that Liza was a good teacher, hopefully making use of the phenomenon that computers are perceived as social actors to motivate the users. (Nass et al. 1995, p. 224)



```
Liza E.: And what's the probability for this answer to be correct, what do you think? 100%? 80%? 50%?
You: 99%
Liza E.: You are probably right with this evaluation.
Liza E.: You would be could be a good scientist! Or at least a good assistant to Mark.
You: Thank you -)
Liza E.: You're welcome.
Liza E.: Are you interested in my view on this question?
You: No, I think I got it.
Liza E.: Okay.
Liza E.: Allow me to go on with the questions.
```

Figure 10: *The agent asks about confidence*

4.2 Parsing

The parsing module of the agent is highly specialized for its core purpose: to assess whether the specific questions asked by the agent were answered correctly. Every input is parsed with regard to the context of the message, and more than 80 different context types are distinguished. Knowing the specific context, the agent can be relatively sure that a certain keyword translates to a right or wrong answer to the question asked, whereas without the context, the agent would be clueless because of its rather simple parsing system. This approach reduces the complexity of the parsing a lot and can be seen as a shortcut to achieve more human-like results with simple methods. The input is sorted into one or more categories which are returned as the parsing result. Besides appearing as an inexperienced agent, the only measure against annoyance that can be taken is to prevent misunderstandings and errors in parsing³¹. Therefore, the parsing algorithm was tested a lot, taking a considerable part of the developing time. Naturally, it is not as good as a real human would be.

³⁰Requirement 4: Providing feedback

³¹Requirement 10: avoiding frustration

4.3 Finding Tags

To assess the meaning of the user's input, the parser searches for tags and finds out which of them occur in the input and if they occur in a positive or negative way. This is done with the help of a list of a list of negative phrases, like "not", "wouldn't" or "anything but", that are put in relation to the tag. If no negative phrase can be found in relation to the tag, it is considered to be a positive occurrence, if one can be found a negative one, if two can be found it is a positive again. This method is invoked by most of the more general parsing methods and provides a core functionality for understanding the input.

"I think both outcomes are **equally** likely to occur."
correct tag "equal" found (not negated)
answer is evaluated as correct.

"I **don't** think the conclusion **follows** from the premises."
incorrect tag "follows" found (negated by "don't")
answer is evaluated as correct.

"I **doubt** that it could be **anything besides** the **second** one."
incorrect tag "second" found (two negations)
answer is evaluated as incorrect

4.3.1 General Parsing

Every input will be passed to a general parsing routine. This routine finds compliments, criticism, expressions of boredom, affection, aversion, uncertainty or excitement, asks for hints, explanations or more time, and some other general statements. The findings are included in the parsing results and used by the main control routine to react to those kinds of input that were not specifically asked for, but will often occur in a conversation with a human. Nearly every situation is handled by a specialized method of the parser. To give two examples: When the agent greets the user, the next input by the user is evaluated from this point of view, looking specifically for typical phrases that occur in greetings like user introducing themselves by their name. When a yes/no question is asked, a specialized methods looks out for phrases of agreement or disagreement.

4.3.2 Situation-Specific Parsing

The arguably most important part of the parser is the one that is responsible for understanding the user's answer to the stories that were asked. To give useful feedback, it is crucial to correctly identify correct and incorrect solutions. For questions that require a decision or answer with certain terms, the user input will be scanned for every tag that is associated with the question. If a correct tag appears in a positive (not negated) way or an incorrect tag appears while being negated, this is evaluated as indicating a correct answer. If the opposite occurs - an incorrect tag appearing in a positive way or a correct one in a negated way - this suggests the answer was incorrect. If no indication for either side, or indications for both sides appear, the parser has no clear result and returns this information to the main control unit.

Many of the questions demand an assessment of probabilities of the user. In this case, the answer is not evaluated using tags, but with a specific method to parse probabilistic statements. At first, numbers and numerals that occur in correct combination with phrases like "percent", "to" or "out of" are found. Combinations like "20 to 40 percent"

are interpreted as their mean value to get one single value. If this returns no results, the parser looks for phrases that describe probabilistic concepts, like “very probable”, “not obvious” or “absolutely sure” and their combination. Those are converted into an appropriate percentage estimate, which is passed back to the control unit to ask the user if this percentage is in line with his original statement or not, offering the option to give another answer. When a concrete percentage is finally obtained, it is matched against the correct answer for the question, and if relative and absolute difference of the two values are within certain bounds, the answer is parsed to be correct, otherwise to be incorrect.

The rest of the parser consists of specialized methods that use tags to interpret the input for certain situations as already described.

4.4 Architecture

Liza consists of several units, as can be seen in Figure 12.

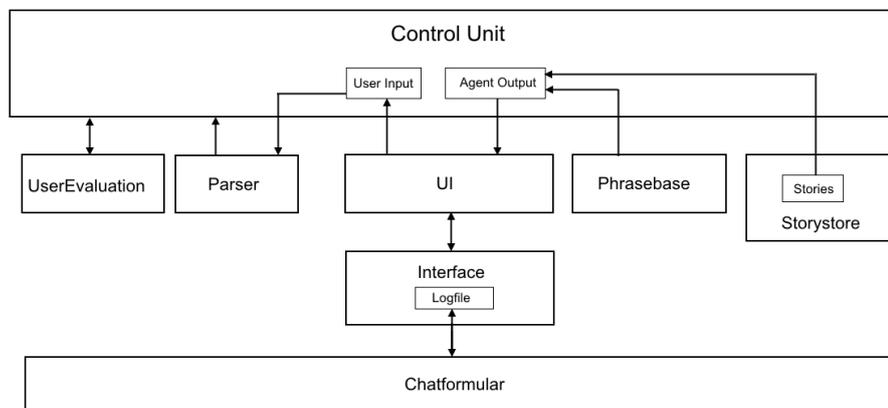


Figure 11: *Relations of the agent's components*

The Control Unit is responsible for the general management of the dialogue - it keeps track of the state of the conversation, decides what to do next and interacts with all the other components. The output for the user is sent to the UI, which interacted with a graphical user interface before the agent was put online, and is now connected to an interface for the web application. User input goes in the opposite direction, being received by the Control Unit from the UI. It is then passed on to the Parser, which returns to the ControlUnit the content that could be extracted. The ControlUnit gets its statements and comments mainly from the Phrasebase, and the stories are retrieved from the Storystore, which holds a collection of all stories. A story is a task that is to be presented to the user, complete with descriptions, questions, reactions, explanation and everything else that is necessary. An example story is included in the appendix. While the ControlUnit is executing the stories, correct and incorrect responses are stored by the UserEvaluation. An overview about the relations between the components can also be seen in figure 11. The separate modules will be described

in detail in the following sections, but only the important features are outlined and many minor details are excluded in favour of a better general overview.

4.4.1 Control Unit

The Control Unit is the heart of the agent and interacts with the other parts. It controls the state the program is in, and initiates every dialogue. In the beginning of the dialogue, it creates a list of exercise questions retrieved from the **Storystore** and uses it to guide the user through the learning process, as is described in section 4.1.2. The Control Unit is the main executive for deciding what the agent does. To put it another way, it 'is' the agent. Output that the agent wants to tell the user is sent by the ControlUnit to the `tell()` method of the UI. The UI's `listen()` method, on the other hand, is used to receive the user's input.

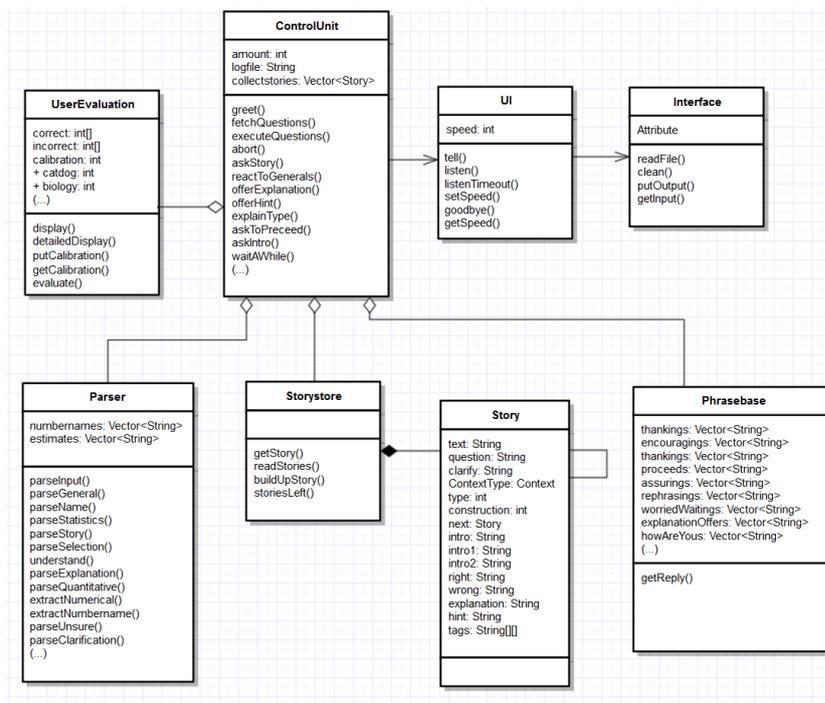


Figure 12: *simplified architecture of the agent*

The ControlUnit also holds the **Parser**, the **Phrasebase** and the **UserEvaluation**. It sends user input in form of the raw string, together with information on the context and the story it belongs to, to the parser's `parseInput()` method and decides how to go on based on the content interpretation the parser delivers back in form of an **InputType** (enum). It requests phrases for various different occasions as Strings from the **Phrasebase**, and it uses the **UserEvaluation** to store information about which questions the user did right or wrong (via the `evaluate()` method) and causes it to submit a complete feedback of the user's performance via the `display()` method. In this manner, the Control Unit conducts the whole teaching process.

4.4.2 UI and GUI / Interface

The UI is responsible for handling input and output via the `tell()` and `listen()` / `listen_timeout()` methods invoked by the `Control Unit`. It also stores the current speed the agents uses for ‘talking’. The UI has one method for telling something to users, but two methods to receive input (or ‘listen to them’): One that is used to wait forever for a reply, only to be interrupted for a general timeout of the whole program after roughly 30 minutes, and one with a definite timeout set. The latter one is used whenever the agent is intended to wait only for a while before inquiring about the user’s well-being, offering hints, making comments and so on. In fact, the timeout-bound listening method is used most of the time, while the other one is reserved for last-chance questions like making sure if the user really wants to quit.

The UI can either communicate with the `GUI`, which was used in the beginning of the development and which provides a local chat window for easy communication, or with the `Interface`, a class which is used to connect the agent to the web application. The `Interface` receives output from and provides user input to the UI, and interacts with a logfile by simply reading and writing and some basic analysis, e.g. inserting a note that Liza is ‘typing’ right now, or that the agent has disconnected.

4.4.3 Stories and Storystore

The ‘stories’ are the questions that are asked by the agent. There are 66 different stories in the current version of the agent, with well over 13000 words in total. Each story contains a text describing the premises and basic situation, a concrete question and a clarifying reformulated question, a category, a set of correct and incorrect tags (signifying correct or incorrect answers), a hint, an explanation and reactions to correct and incorrect answers. Most of the stories also have an introductory bit of chat with a question leading to the topic, and different reactions for the introduction.

The `ControlUnit` picks stories from the `Storystore` by topic using `getStory()` and can also assess a single one via the method `getSpecificStory()`. But the `Storystore` does not only provide the stories, it also offers methods to alter the stories according to the knowledge that was gained during the conversation with the user.

4.4.4 Phrasebase

The `Phrasebase` is the name for the module that stores more than fifty collections of each five to twenty interchangeable phrases for every situation. Its sole purpose is to answer requests for phrases (for example for variants of approval) with a randomly selected phrase from the appropriate collection (in this case for example “Ok”, “Fine”, “Alright then” etc.). The `ControlUnit` interacts with the `Phrasebase` by requesting a phrase for a certain situation, e.g. a greeting, via the `getReply()` method. The `Phrasebase` then randomly picks one of the phrases of the greetings-category and returns it as a string.

While this is technically simple, it plays an important role in avoiding repetitions, thereby letting the agent appear human. If the name of the user is revealed to the

agent, a portion of the phrases in the `Phrasebase` is updated to contain the name via `setName()`, which allows for a more direct approach in talking to the user.

4.4.5 Parser

The `ControlUnit` sends user input in form of a `String` to the `Parser` to ‘understand’ what the input means, together with the context it was given in (via `ContextType`, an enum) and, when it was an answer to a specific story that was asked, the story itself. The `Parser` then interprets the input with regard to the context and story, and returns a collection of his findings (also enums) back to the `ControlUnit`. Those findings for example include ‘`NO_IDEA`’ for cases where the parser was unable to make sense of the input, ‘`CORRECT`’ for a fitting solution when a question about a certain story was asked, ‘`SLOWER`’ when the user asked the Agent to reduce its speed, or ‘`SELFCRITIC`’ when the user said something along the lines of “I am stupid.”

The parsing module of the agent is highly specialized for its purpose. Every input is parsed with regard to the context of the message, as was explained before. There is a `ContextType` for every dialogue situation that is distinguished by the agent, e.g. ‘`GREETING`’ for the situation where the agent just greeted the user or ‘`YESNO`’ for a situation when a yes/no question was asked. Also, there is one `ContextType` for each story, to associate the user’s input with the story it belongs to. After all, answers can only be interpreted with respect to the question that they belong to. In sum, more than 80 different context types are distinguished. This approach reduces the complexity of the parsing a lot and can be seen as a shortcut to achieve more human-like results with simple methods. A more detailed description of the parsing algorithm can be found below in the conceptual design section.

4.5 Implementation

Liza is textual, with the small exception of graphs, pictures and diagrams shown by the agent to illustrate explanations. There were no indications of inappropriately high expectations from the users, as they stuck to the tasks Liza gave them and followed her guidance easily. The agent was developed as a java program, without any additional tools. Although it first seemed useful, AIML was not used in creating the agent, because its constructs seemed not flexible enough. The agent has a mixed-initiative approach and also does not always immediately react with a response to a given input, but displays very different behaviour depending on the context, so a more variable approach was needed. In general, AIML offered no major benefits compared to coding the behaviour directly (Fonte et al. 2009). There are also language analysis tools, like *FreeLing* (*FreeLing Home Page* 2016), a c++ library used for example in TQ-Bot (Fonte et al. 2009), which offers a variety of language parsing options like morphological analysis, semantic role labelling, parsing etc., just to give one example. Those tools are not needed for the highly specialized approach that Liza’s parsing takes. The agent is not required to offer a totally free conversation, but instead mainly needs to find out if the answer to a certain question falls in one or the other category. This can be

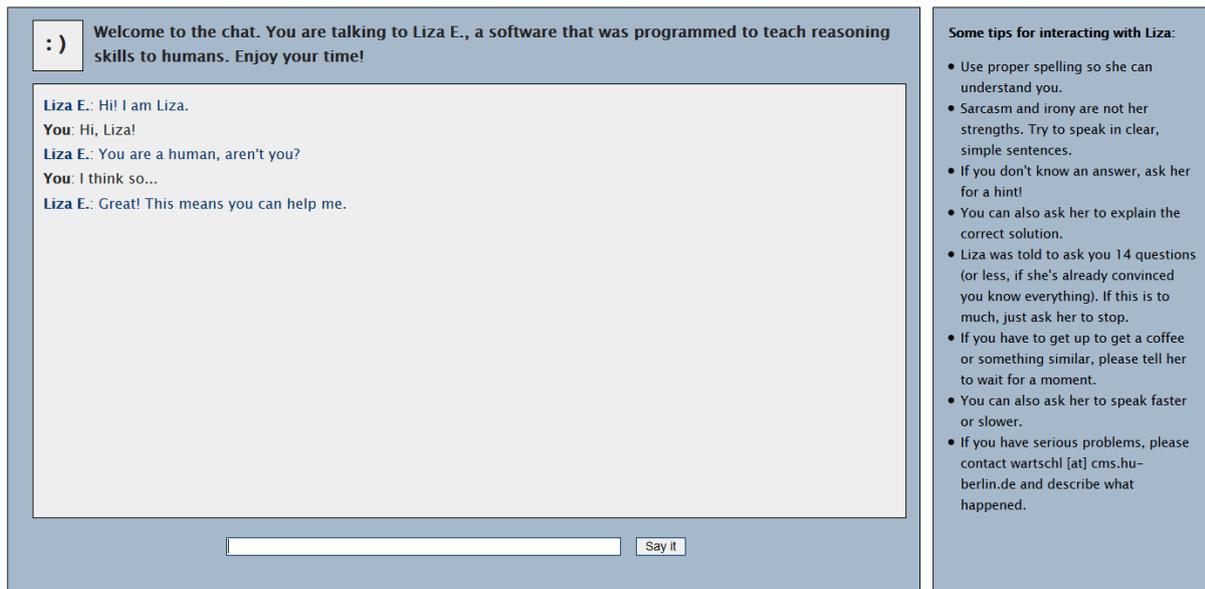


Figure 13: *web chat with 'Liza'*

done very efficiently with the somewhat elaborate pattern matching algorithm that was presented before in section 4.2. Therefore, no advanced language parsing tools were used.

The web interface needed for the survey was created using mostly php and html forms on a privately set up server. For every user, a new instance of 'Liza' is executed. The web page and the java program running made use of a logfile to keep track of the progress of the chat, with both sides regularly checking the file for new messages. In this manner, mixed-initiative dialogue was implemented. Next to the chat window, a list with some tips for interacting was presented to help users to handle the agent, as can be seen in figure 13.

The option for leaving the chat only appeared after Liza disconnected, which could be triggered by either completing the full conversation and her saying good-bye to the user, or by asking her to stop and insisting on a premature end.

5 Teaching Content

The area of human biases and errors is a vast one, and not every failure of the human mind has been researched in detail. This work focuses on seven different especially well researched heuristics or biases that frequently lead to irrational thoughts and decisions. They are associated with understanding chance and probabilities, evaluating evidence, making economically rational decisions and being able to understand logical conclusions. The general idea or hope is that the agent will increase the understanding of those errors to help people avoid making errors, to raise the awareness for this

kind of fallacy, and to help users adjust their calibration for how sure they can be of their solutions. While learning about biases is actually nothing else than improving mindware, the second and third goals (raising awareness and improving calibration) would make people generally more likely to actively use their system 2 because they are more aware of their susceptibility to errors. Both of the approaches, improving mindware as well as using system 2 more, are viable ways of increasing reasoning skills.

5.1 Topics Taught by the Agent

5.1.1 Bayesian Reasoning

Probabilistic thinking and dealing with conditional probabilities does not come easy to most people, especially when different probabilities have to be combined according to *Bayes' Theorem* to access the correct result. The ability to do so is tested with problems like the following (Stanovich 2009b, p. 37):

“Imagine that XYZ syndrome is a serious condition that affects one person in 1,000. Imagine also that the test to diagnose the disease always indicates correctly that a person who has the XYZ virus actually has it. Finally, suppose that this test occasionally misidentifies a healthy individual as having XYZ. The test has a false-positive result of 5 percent, meaning that the test wrongly indicates that the XYZ virus is present in 5 percent of the cases where the person does not have the virus. What is the probability that a person has the XYZ virus, given that the person’s test result was positive?”

The background of the task above is a medical one because physicians have to deal on a daily basis with subjective probabilities the degree of confidence for a patient’s condition. Reacting correctly to new evidence, such as a positive or negative test result, is crucial for their patient’s well-being. Only 1 of 60 medical professionals and students of the Harvard Medical School were able to produce the correct result (2%) for an equivalent test problem (Casscells et al. 1978, p. 999). Their answers varied a lot, from 95% to 2%, and caused the authors of the study to conclude that “formal decision analysis was almost entirely unknown and even common-sense reasoning about the interpretation of laboratory data was uncommon among them.” (Casscells et al. 1978, p. 999-100) Further investigations have confirmed this result (Eddy 1982; Berwick et al. 1981).

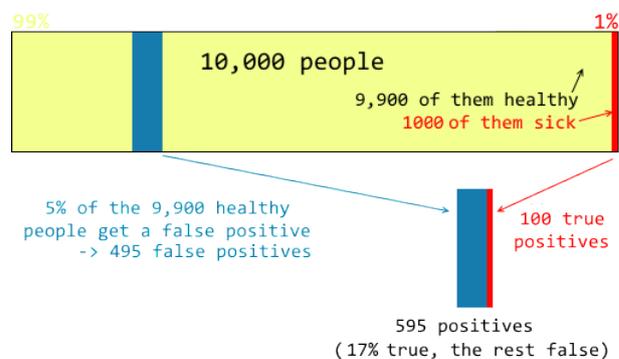


Figure 14: An image used by the agent to explain frequency formats for a similar problem

Assume the probability that a patient is ill is called $P(A)$ and the probability for a positive result is $P(B)$. The most common mistake people make is to take the probability that the test will show a positive result, given that the patient is ill, $P(B|A)$, to be nearly the same as the probability for the patient to be ill, given that he or she has a positive result, which is $P(A|B)$. The base rate or prior probability $P(A)$, in the example above 0.1%, is widely ignored, as people focus on the more specific true positive rate of the test, which seems intuitively to be more important to the result, although all three given facts have to be considered to achieve the correct value. This error is called *base rate neglect*, and it occurs not only in bayesian reasoning problems, but in a variety of situations (Tversky et al. 1974, p. 1125).

Even people who have had extensive statistical training predict the outcome that seems most representative for $P(B|A)$ when they use their intuition to solve problems (Tversky et al. 1974). But it has been shown that although teaching bayesian reasoning is by far not simple, there are useful methods to teach the correct approach. The key is to translate probabilistic problem into a frequency format. Instead of thinking about various probabilities, people are taught to imagine a certain group of individual cases (in the above case maybe 10000 people), and divide them in subgroups according to the probability ratios, e.g. $P(A \text{ and } B)$, $P(A \text{ and not } B)$, $P(\text{not } A \text{ and } B)$ and $P(\text{not } A \text{ and not } B)$. In this format, the problems are much easier to solve, making the training in frequency formats the most successful approach for teaching this kind of reasoning, although it is still a very difficult subject (Hoffrage et al. 1998; Larrick 2004, p. 325). This strategy is used by the agent Liza to teach the correct solution. It tells the user to take a large population and divide it into subgroups, e.g. determining for 10000 people how many are ill, how many are healthy, what share of each of those groups will get a positive result (and the concrete number), and add them up to see how many people in total will have a positive result. Now the result can be easily determined by dividing the amount of people who are actually ill and get a positive result by the number of *all* people who get a positive result. This concept is emphasized by images to help visualize the approach. Figure 14 shows one such example, using slightly different numbers than the quoted problem by Stanovich.

5.1.2 Law of Large Numbers / Regression to the mean

Another key part of reasoning about uncertainties is dealing with sample sizes and the effects tied to them. The Law of Large Numbers states that when performing many trials of random, independent events, the average of the result will be much likely to be close to the expected value than for only a few samples. The related phenomenon of the regression towards the mean, first described by Francis Galton in the late 19th century, states that if an extreme result is achieved in a first trial of an independent random event, the second event will be extremely likely to be closer to the average (Tversky et al. 1974, p. 1126). Both laws are important to assess the (lacking) significance of single events or small samples that are influenced by chance.

In general, people tend to ignore the size of the sample when they make decisions (Tversky et al. 1974, p. 1125). The qualities of a single piece of evidence (for example, how impressive or representative it is, if one can relate to it, or the effect size) is of more importance to them than its informational value (e.g. the trustworthiness of the source or the underlying sample size) (Griffin et al.

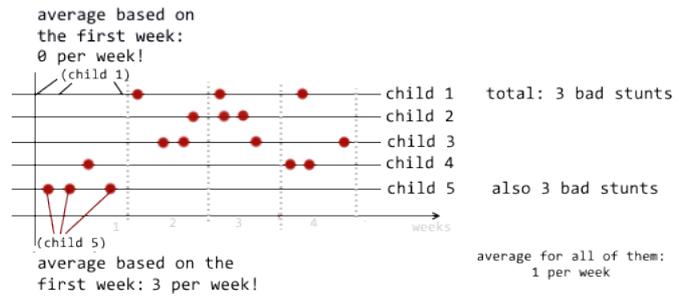


Figure 15: *Graphical representation used by Liza for a problem about the regression to the mean of student's behaviour*

1992, p. 411; Stanovich et al. 2000). The regression towards the mean is also widely disregarded, which can lead to serious misconceptions: teachers in flight training notice that after particularly bad performance by a trainee at landing a machine, followed by severe criticism from the trainer, the next landing is typically way better. They assume that strong criticism as a punishment improved the performance of the trainees. Of course, for praise after exceptionally good landings, they observed the inverted effect, concluding that positive messages did not motivate the trainees to become better. Both conclusions are caused by the regression towards the mean and are dangerously inconsistent with psychological research on praise and punishment (Tversky et al. 1974, p. 1127).

Fortunately, people can be trained to understand and apply the correct statistical principles in such cases. A training on the general concept of the Law of Large numbers already causes people to apply it in ordinary problems in their life. A training that combines abstract rules with tangible examples has proven to be more effective than using just abstract concepts as well as giving only examples (Fong et al. 1986).

Of course, this is exactly the approach the presented agent will take to teach its users. As it does with each of the topics, the importance of sample size and regression to the mean is explained as a general concept and illustrated with examples. Different graphical representations are used to help the user grasp the concept, as shown in figure 15.

5.1.3 Gambler's Fallacy

The Gambler's fallacy is the misconception that independent random events have some kind of memory of the last results and are obliged to even out, influencing the odds for the next outcome after a series that is perceived as a 'streak'. After not rolling a six for ten times with a six-sided die, people tend to believe that the six is now 'due' and more likely to occur than in the beginning. As Kahnemann and Tversky (1974) put it, "Chance is commonly viewed as a self-correcting process in which a deviation in one direction induces a deviation in the opposite direction to restore the equilibrium. In fact, deviations are not *corrected* as a chance process unfolds, they are merely diluted."

A typical gambler’s fallacy problem that is used in research on the topic would be the following (Toplak et al. 2011):

“Imagine that we are tossing a fair coin (a coin that has a 50/50 chance of coming up heads or tails) and it has just come up heads 5 times in a row. For the 6th toss do you think that:

- a. It is more likely that tails will come up than heads.
- b. It is more likely that heads will come up than tails.
- c. Heads and tails are equally probable on the sixth toss.”

The fallacy is so common that only 41% of middle adolescents presented problems involving it gave the correct normative response (Klaczynski 2001, p. 852). It seems to be quite difficult to teach people to avoid the gambler’s fallacy, as early tries resulted in no significant differences in display of the bias between people who were taught about it and a control group (Beach et al. 1967, p. 281). A possible explanation for why people persistently view chance that way was provided by Gestalt psychology, assuming that people have to decide whether they want to view the next event as part of the group of past events or as a single entity. The former makes people think about the unlikelihood of heads occurring six times in a row (even though 5 of them occurred already and are ‘fixed’), leading to the gambler’s fallacy, while the latter makes them focus purely on the attributes of the next event. The theory has not been tested thoroughly, but it suggests that a possible approach would be to teach people to choose the correct perspective on events and explicitly avoid viewing independent random events as part of a group (Roney et al. 2003). The agent uses this approach to teach the concept to its users.

5.1.4 Wason’s Selection Task

Besides reasoning about statistical rules, there are also widespread biases in the domain of logical reasoning. P. C. Wason designed the selection task that tests the understanding of conditional rules and the ability of the subject to not only jump to the first assumptions about a rule that comes to mind, but to see through it completely and falsify it. In its classical form, the task is presented as following:

Four cards are presented to the test person, one showing a vowel, one a consonant, one an even number and one an odd number. See figure 16 for an illustration. The participants are informed that each card has a letter on one side and a number on the other. The following rule is told to them by the experimenter: “If there is a vowel on one side of the card, then there is an even number on the other side,” and the subjects are told to identify all the cards that are needed to find out if the experimenter was lying to them, but no more (Wason 1968, p. 273;



Figure 16: *illustration used by the agent when presenting the selection task*

Klaczynski 2001, p. 848).

In this abstract form, the percentage of people getting it right (by choosing the vowel and the odd number) is below 10%. Most people chose the vowel and the even number, or insist on turning all the cards, even after several trials (Cheng et al. 1986, p. 295). In a study testing mathematics and history students (Inglis et al. 2004), 8% of the history students solved the task correctly, and 29% of the maths students managed to do so. Maths staff members outperformed them with 43% correct results. The authors of the study suggest that the reason for this could be the commonness of error checking in the daily work of a mathematician, which leads them to falsify their own initial thoughts more than the historians do. A similar explanation for the general failure of people to solve this task is given by Stanovich, who argues that they tend to confirm rules rather than trying to falsify them, and therefore focus on the two cases mentioned instead of thinking about possible falsifications (Stanovich 2009b, p. 38). Wason himself suspects that the subjects did not consider the task to be particularly difficult, preventing them from grasping the full meaning of the rule. But he also speculates that they might simply not have the ability to carry out formal operations on variables by forming propositions about them [p.280-281] (Wason 1968). When reformulating the condition with a real-life example based on rules and permissions such as minimal drinking age or passport requirements, about 60% of the subjects give the correct response (Cheng et al. 1986, p. 294-186). But since the agent described here is designed to improve general reasoning, which also includes falsifying own beliefs, the agent does not teach the subjects to view the logical rules as permissions or prohibitions. Instead, it focuses on highlighting the importance of considering *all* the cases, thinking calmly about the possibilities, and questioning the first impression that comes to mind.

5.1.5 Covariance Detection

	Improvement	No improvement
Treatment	200	100
No Treatment	75	25

Another example of a situation in which people do not try to falsify a hypothesis is assessed with a covariation detection task, also called contingency detection. In this case, the hypothesis is an assumed connection between a certain action (e.g. a medical treatment) and an outcome (e.g. patients recovering). A comparison between the results of the action in question on the one hand, and the results of a control group on the other hand is presented. The subjects have to judge whether the action was efficient. This has many applications in everyday life (e.g. “is gun control responsible for a decrease in crime?”) (Stanovich et al. 1998, p. 170). To falsify the hypothesis of the treatment’s effectiveness, subjects have to examine the no treatment condition and compare its results to the treatment condition. The ability to look out for a confounding variable that is responsible for the improvements is taken as an indication about the subject’s ability to reason about confounding variables in general (Toplak et al. 2011, p. 1286).

To solve the test correctly, it is necessary that the subject compared the ratios of improvement and no improvement cases. To apply this conditional probability rule, no complicated calculations were needed. The first impulse, however, is just to look at the large number of cases in which treatment and improvement occur (in this example, the 200 cases), without thinking about which conclusions can actually be derived from that (and which can not) (Stanovich 2009b; Klaczynski 2001).

Liza will warn the user about jumping to conclusions about the effect of the treatment. Instead, it tells them to calculate the ratios for both cases and to compare them. This is accompanied by a general warning about the dangers of focusing on the first impression that comes to the mind. The agent uses cases in which the treatment was effective as well as cases in which it was totally ineffective. Again, images and tables (as shown in figure 17) are used to illustrate the case.

	growing well	growing ok
new fertilizer	9	3
old way	14	14

Figure 17: *Illustration used by the agent when presenting the covariation task*

5.1.6 Sunk Cost Fallacy

Another typical fallacy that diminishes people’s ability to come to reasonable decisions is the sunk cost fallacy. It occurs when people fail to acknowledge that money or time that has been spent in the past and *cannot be retrieved* should not influence future decisions. The cost that was spent is *sunk* when nothing can be done about it anymore. Sunk costs play no role in rational decisions (although one can of course use knowledge gained from former errors) (Friedman et al. 2007, p. 80).

While the literature on the sunk cost fallacy is vast and sometimes contradicting, two main reasons for the occurrence of the fallacy have been established: cognitive dissonance, that makes people feel very uncomfortable with acknowledging a past mistake, leading to further investment of time or money in a lost cause, and excessive loss aversion, that keeps the hope alive that the sunk costs could be eventually recovered, leading to further commitment or investment to the same cause (Friedman et al. 2007, p. 87).

Typically, the sunk cost fallacy is assessed with questions like the following (Larrick et al. 1990, p. 363; Thaler 1980):

“[I]magine you have bought \$15 tickets for a basketball game weeks in advance, but on the day of the game it is snowing and the star player you wanted to see is sick and will not play. The game is no longer of much interest to you. Do you decide to go to see the game or stay home and forget the money you have spent?”

The ability to overcome regret about past decisions and reason only based on what is relevant now can be learned. Non-economists are less likely to apply coherent

cost-benefit rules than economists are (Larrick et al. 1990, p. 365). In a relatively brief training sessions, the performance of individuals can be improved not only in the trained domain, but also for other areas where they have to generalize their knowledge, even swapping financial with other costs (and vice versa). The subjects were able to apply their training even a month later. (Larrick et al. 1990, p. 365-367).

The agent points out the fact that the sunk costs have been spent, no matter what decision you take, using simple charts to visualize the subject matter. It encourages the user to always try to make the best out of the situation they have.

5.1.7 Belief Bias in Syllogistic Reasoning

People suffer from belief bias when they evaluate the validity of logical arguments based on beliefs about the world that are not stated in the arguments they evaluate. Arguments that seem natural or believable are viewed as being more *logically consistent* than those that have unusual or unfamiliar contents (Evans et al. 1983, p. 302). A typical example is: (Stanovich et al. 2000, p. 647)

“All mammals walk. Whales are mammals. Conclusion: Whales walk.”

The question posed to the subjects is: Does the conclusion logically follow from the premises? In this example, it does, but (because the first premise is not true for the world we live in), the conclusion seems absurd and implausible. The opposite problem arises with a question like the following one: (Markovits et al. 1989, p. 12)

“Premise 1: All flowers have petals. Premise 2: Roses have petals. Conclusion: Roses are flowers.”

Here, the conclusion seems plausible, but is invalid. When subjects have to deal with beliefs from their everyday life that conflict with the validity of the conclusion, the percentage of people who fail at the reasoning task goes up as high as 92% (compared to only 27% for cases in which believability and validity correspond) (Evans et al. 1983, p. 298). When presented with a more abstract framing of the task in which beliefs have a less strong impact, people respond with more normatively correct responses (Markovits et al. 1989, p. 14). Since belief bias alone is not the only reason for errors people make in those tasks (Markovits et al. 1989, p. 15), the performance on syllogisms in which what is plausible corresponds to what is valid (plausible and valid *or* implausible and invalid) must be compared to the performance on tasks where plausibility and validity are opposed to each other, as they are in the statements given above. The ability to solve the tasks correctly where plausibility and logic conflict is generally considered a quality, as it requires putting away prior knowledge and assumptions and reason independently in a situation with new

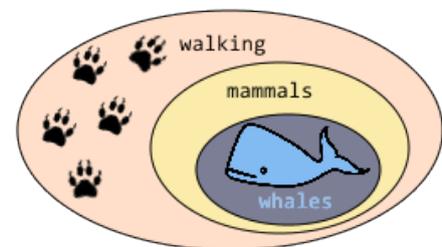


Figure 18: *Illustration used by the agent to show the validity of a conclusion*

premises (Toplak et al. 2014, p. 153).

Liza explains to people that the ability to abstract from the content of the statements to a general rule is important and shows Venn diagrams to illustrate the relationship between the statements, as is illustrated in figure 18. It demonstrates the user how an abstract variant of the phrase in question would work, replacing 'whales' or 'mammals' with 'A' or 'B', and points out that only if it still seems valid, the conclusion can be logically sound.

5.2 Relation between the reasoning tasks

While many people give fallacious and non-normative answers to the tasks mentioned above, there are always some who are able to solve them correctly. It seems that there is some quality in humans that accounts for a general ability to solve those reasoning tasks - the ability to reason. In four studies with 954 participants in total, Stanovich and West investigated the correlation between various reasoning tasks. In the first, they let their subjects perform syllogistic reasoning tasks and selection tasks (amongst others) and found a significant correlation between the two tasks (.363). In another experiment, significant correlations with the covariation detection task, the syllogistic reasoning task, the selection task and some others were established. Furthermore, studies suggest that there is internal consistency between various decision making tasks like the Sunk Cost Fallacy, suggesting that the ability to solve them normatively is not due to random occurrence or absence of errors, but a distinctive quality of a person (Bruine de Bruin et al. 2007). Klaczynski investigated Wason's selection task, contingency detection, statistical reasoning about sample sizes, the gambler's fallacy, and other biases. He found that there were statistically significant correlations between the normative as well as the non-normative solutions to the tasks (Klaczynski 2001). Toplak et. al examined the base rate fallacy, bayesian reasoning, sample size problems, regression to the mean, the Gambler's Fallacy, covariation/conjunction problems, the Sunk Cost Fallacy and some other biases and found a high correlation with the cognitive reflection test, independent from intelligence (Toplak et al. 2011). In general, the heuristics and biases tasks are connected with each other, and evidence suggests that the individual differences in the ability to solve them can partly be reduced to a general reasoning ability. For Liza, the heuristics and biases tasks that were included were chosen according to the following principle: a connection with other tasks should already be established, they should be teachable or trainable, and they should be assessable with only one or two questions. By these criteria, the seven already explained tasks emerged.

5.3 CRT

As already hinted in the descriptions of the various biases above, faulty reasoning often occurs when people follow a first, intuitive impulse that leads to an erroneous solution, rather than thinking a problem through. To withstand this impulse, one must stop and question the intuitive response, check it for validity, or search for other ways of solving

the task. The cognitive reflection test (CRT) measures the ability to do exactly this. It has a substantial correlation with cognitive ability, and it has been shown that the cognitive reflection is also an important predictor of people's performance on common biases tasks (Frederick 2005; Toplak et al. 2011). Since its introduction, it has been a popular tool of measuring dysrationalia, probably also because it can be executed with a small set of questions. The following problem is a key component of the cognitive reflection test and illustrates its main principle well (Stanovich 2009a):

“A bat and a ball cost \$1.10. The bat costs \$1.00 more than the ball. How much does the ball cost? _____ cents”

The intuitive answer is '10 cents', but it requires not a great deal of mathematical sophistication to find out that the correct solution is 5 cents. Anyone who thinks about this problem for a moment notices this. But yet, many people give the wrong solution because they do not invoke their ability for cognitive reflection, but rather just follow their intuitive system 1 answer.

All the questions of the CRT follow this principle - the correct answers is not hard to deduce, but is superimposed by an immediately available intuitive response. People who solve the tasks correctly often rethink their first idea (sometimes crossing out their first solution of 10 cents and writing the second, correct one down next to it, but never the other way round), and they consider the task to be harder than those who just go with their intuition (Frederick 2005). The strong intuition that must be overridden with the correct analytical solution goes well with the two systems theory of reasoning (as described in section 2.3.2, *'The Dual Process Theory'*).

Correlations of the performance have been shown with temporal discounting, gamble behaviour, choices consistent with expected-values, avoiding the conjunction fallacy, rational profit maximising strategies, performance calibration and non-superstitious thinking (Toplak et al. 2014, p. 149). An expanded version of the CRT has been found to correlate strongly with heuristics and biases tasks such as belief bias in syllogistic reasoning, selection task, denominator neglect, bias blind spot and temporal discounting (Toplak et al. 2014, p. 159).

The agent will *not* teach anything about the problems that are used to assess cognitive reflection. Instead, it will just focus on highlighting the importance of question first intuitions in general, creating awareness for how easily people can make mistakes, and encourage the users to solve problems analytically. But the CRT scores of the subjects will be measured along with the heuristics and biases scores to see if any effect on them can be found.

5.4 Critical Thinking and Argumentation Skills

There are related fields to the research on reasoning and rationality that could have been included in this work, but are not. Especially critical thinking, which involves thinking skills like openness to new points of view, ability to evaluate arguments, and inferential reasoning, has some overlap with the general idea of rationality described here. But

since it is a very vast topic, it would have inflated the scope of the agent too much. Furthermore, there is already some research on the applicability of computer-supported technologies in teaching this kind of thinking skills (More 2012; Pinkwart et al. 2012).

6 Experimental Evaluation

6.1 Goals

To evaluate the impact of the agent, the effect on the students has to be examined. The following six hypotheses are tested: 1. The participants will do better on the reasoning tasks after they practised with the agent than they did before, in other words, they improve their reasoning skills in those areas. 2. Participants who took part in a non-interactive online course will also improve their reasoning skills. 3. Talking to the agents leads to a stronger improvement than the non-interactive online course. 4. In the second test, the performance of the participants who talked to the agent will be significantly better than the performance of the participants who took the online course.

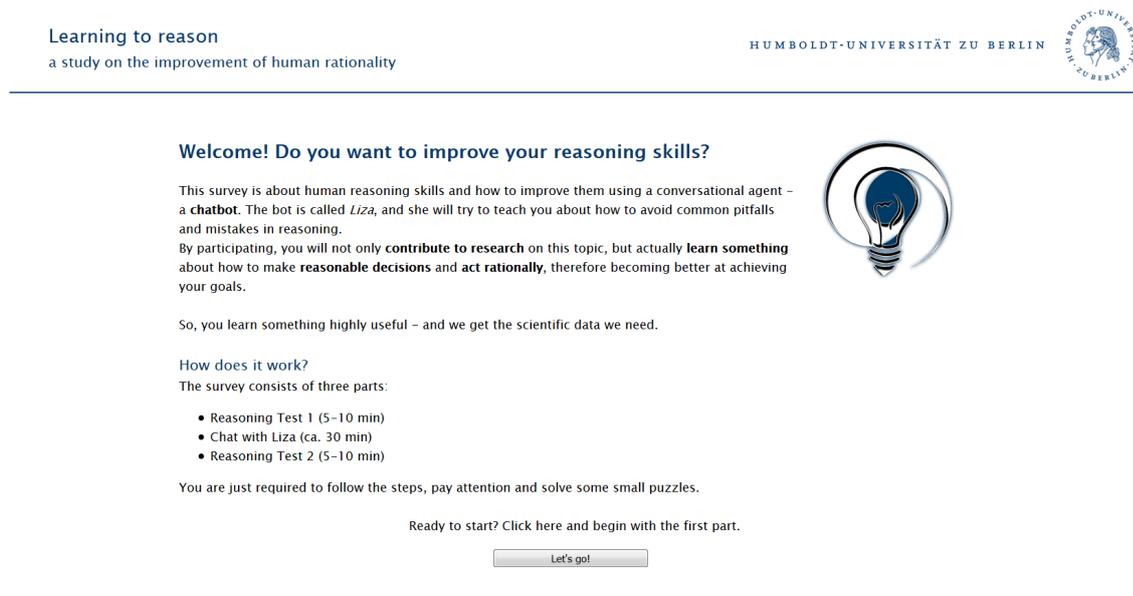


Figure 19: *welcome page for the survey showing the experimental condition*

A fifth hypothesis is tested, too: Working with the agent improves the general readiness of subjects to use their System 2 (thoughtful reasoning) instead of System 1 (intuitive responses), as described in section 2.3.2. This can be measured by the Cognitive Reflection Test which was introduced in section 5.3. Although it seems very improbable that a short intervention of roughly 30 minutes would be enough to produce such a general result, the results will be checked for this, too. The last hypothesis is the following: The results of the training with Liza will still be

present after two weeks, indicating that long-term benefits can be gained from working with such an agent.

6.2 Design and Setup

To find out if the agent actually increases the user’s reasoning skills, the following study design was used: The group of test subjects was randomly distributed in a treatment group and a control group. The treatment group got to talk to the agent, while the control group went through an alternative approach: They read seven short texts, each on every of the biases tasks, and learned about the underlying principle of reasoning. For every participant, a test before the intervention and another one afterwards were administered and the performance for every single task were collected. Having finished those, they were invited to leave their email address to be contacted for a follow-up test one to two weeks later. Therefore, every participant could fill out up to three versions of the test. While the tasks were of the same type, they were not identical. To make sure no systematic bias was introduced through differences between the tests, one version of the test served as the pre test for the first participant, post test for the second, and follow up test for the third and so forth.

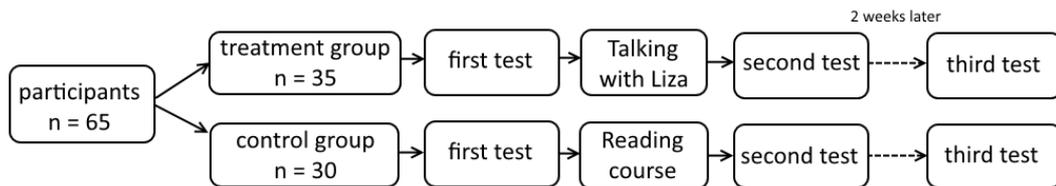


Figure 20: *setup for the experiment*

The improvement of the participants in the treatment group (differences between mean performance for every single task in the first test to the second test) is evaluated, as is the same improvement for the control group. For every single task, the improvements are compared. Furthermore, the difference in performance on the second test is assessed, and it is checked whether the treatment group’s performance is significantly better than the control group’s performance.

The test subjects were allowed to take as much time as they needed. Nevertheless, it was planned to exclude participants who took more than two hours, due to the very high possibility that they got distracted by other activities.

For every participant, some general information was collected: age (younger than 12, 12-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, older than 74), gender (female, male, none/other/not applicable), degree of education (no schooling completed, completed primary education, completed secondary education, completed post-secondary non tertiary education, Bachelor or equivalent, Master or equivalent, Doctorate degree or higher) and level of English (Beginner, Intermediate, Fluent, Native Speaker). Also, participants were asked to specify if one of the following conditions applied to them:

- took psychology or neuroscience courses
- have read ‘Thinking, fast and slow’ by Daniel Kahnemann
- Member of the LessWrong community
- Already familiar with heuristics and biases
- Recently took a Cognitive Reflection Test (CRT)

People who have had psychology or neuroscience courses are very likely to already know about the concepts the agent is talking about, as well as those who are already familiar with heuristics and biases. Members of the online community of ‘LessWrong’ fall in the same category, since this community has a great interest in those topics. The book ‘Thinking, fast and slow’ summarizes the research of Daniel Kahnemann on the topic of reasoning and is one of the best selling books on the topic, reaching a wide audience. Many of the typical questions that are used in research on reasoning and also in this survey are already explained in this book. Finally, the questions of the Cognitive Reflection Test are pointless to assess the cognitive reflection of people if they already know them, therefore those four participants have to be excluded.

Please pick your **age** from the following list:

- younger than 12
- 12-17
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65-74
- older than 74



Please specify your **gender**.

- female
- male
- none/other/not applicable

Figure 21: *extract from the demographical questions*

The instruction used to explain the task to the users was worded as following:

Please solve the following questions by giving the answers that seem most **reasonable, sensible or logical**. It’s perfectly normal that some of the answer may seem obvious (they probably are - there’s no mean trick behind them), and others are harder. Feel free to use a calculator or make notes if it helps you, but please solve the questions **on your own**.

In general, just trust your good reason!

Those instructions were the same for every test condition. The subjects were forced to pick an answer - leaving a question empty would result in a request to answer it - and had to decide for an option even if they were not totally sure. This was done to assess even slight preferences to one pick, despite subjects being not totally sure about their reply.

6.2.1 Origin of Texts and Tasks



Assume the following two premises are true: *All cats have fur. Sphynx cats don't have fur.* Can you logically conclude from this that *sphynx cats are no cats*?

- invalid (conclusion does **not** follow logically from the premises)
- valid (conclusion follows logically from the premises)

A startup developed a new health app to remind user to do their workout regularly. It is tested on some volunteers for one month, with the following results:

	kept up workout	quit workout
used app	20	8
did not use app	5	2



As you can see, 20 people used the app and managed to keep up their exercises; 8 used the app but quit within one month; 5 were not using the app and stuck to the exercises, and 2 people did not use it and quit. Was the app effective in motivating people to continue their workout?

- yes (effective)
- no (not effective)

Figure 22: *extract from the questions*

The questions were, as far as possible, taken from already established studies in the field. Every test consisted of eleven questions that were mixed in order. Since there were three different versions of the test, 33 questions in total were needed. For each test, there was one selection task, taken from Cheng et al. (1986, p. 301) or Wason (1968, p. 273), one question about regression towards the mean (all from Fong et al. 1986, p. 279), one about the gambler's fallacy (basic example from Chiesi et al. 2011, p. 346 and two newly created variants added), one about the sunk cost fallacy (Larrick et al. 1990; Arkes et al. 1985, p. 363), one bayesian reasoning problem (Stanovich 2009b; Sedlmeier et al. 2001, p. 37 and one newly created), one syllogistic reasoning problem in which belief and logic were coherent (self fabricated) and one in which belief and logic pointed in different directions (taken from Markovits et al. 1989, p. 12), one covariation detection task from Klaczynski (2001), with and two newly created variants added, and finally, for each test, three items for the cognitive reflection test that were taken from Toplak et al. (2014, p. 149), and expanded so that 9 different questions could be asked.

The two different versions of the syllogistic reasoning task were included to assess if subjects did actually follow their logical reasoning - checking 'valid' if and only if the conclusion followed from the premises, no matter if it seemed plausible or absurd - or if they simply tried to trick the system by going always for the 'implausible' answer, suspecting those would have to be the correct ones.

The texts for the non-interactive online course were carefully selected as to provide realistic examples of actual teaching material that would make for a serious comparison to the agent’s work. The explanations on bayesian reasoning and on covariation detection were quoted from Keith E. Stanovich’s award winning book “What intelligence tests miss” (Stanovich 2009b). The gambler’s fallacy was explained using quotes from a textbook used in introductory courses on probability and induction (Hacking 2001). An online accessible introductory textbook (created and used by members of University of Houston - Clear Lake and University of Houston - Downtown) on statistics provided the explanations for the regression towards the mean (Lane et al. 2016; Tunstall 2016). An online textbook of the University on Albany on the topic of management accounting was quoted for the sunk cost fallacy (Caplan 2006). ‘Thought and Knowledge’, a best-selling textbook on critical thinking, was used for explaining the belief bias in syllogistic reasoning (Halpern 2002), and in lack of a specific textbook on the topic, Wason’s Selection task was taught using the explanations Cheng et al. (1986) used for teaching their test subjects.

6.3 Conduction of the Evaluation Study

Over the course of six days, participants were invited to take part in the experiment. During this time, the paid subjects were acquired and acquaintances were asked to take part in the study. The link to the study was posted on several different online platforms, including facebook and reddit. An email was sent to every participant two weeks after they had completed the first part of the survey to invite them to proceed with the second part (test 3 as shown in figure 19).

No major difficulties occurred during the conduction of the study. One of the first participants pointed out an error in the display of a single image file, which could be corrected without stopping the agent.

For every topic that was assessed with a task, the mean performance before and after the treatment was calculated. For every task, a correct solution was evaluated to have the value ‘1’, an incorrect answer was assigned the value ‘0’. For the CRT which consisted of three questions, the value was determined by adding the number of correctly solved questions (between 0 and 3).

6.3.1 Characteristics of the Participants

The participants were acquired using different methods: 25 were paid subjects from clickworkers, a website that offers task crowdsourcing. 14 were acquired with \$3 amazon.com gift cards by posting the offer in various online platforms. The other were either personal acquaintances and were directly asked to participate, or reacted to posts and invitations on various online platforms, and were not paid.

In total, 74 participants of them finished the whole survey. Only nine participants completed the third test two weeks later. They all came from the treatment group, maybe because they agent seemed more interesting to them than the texts were to the

control group members.

Four participants did not fully complete the dialogue with the agent, and quit prematurely, two of them stating technical difficulties due to accessing the website from their smartphones in the feedback form, the others providing no reason. Those participants are excluded from the evaluation. No participant took more than two hours, so nobody had to be excluded because of this. Five of the participants already had an exhaustive knowledge prior to the study by having read Daniel Kahneman’s ‘Thinking, Fast and Slow’ or being knowledgeable about heuristics and biases. They were excluded from the evaluation since experts on the topic are not the intended audience for the agent. That leaves 65 test subjects, of whom 35 talked to the agent (treatment group) and 30 took part in the online course (control group).

The age, gender distribution, level of English and education of the participants can be seen in table 1. One person (1.3%) answered the question about their gender with ‘none/other/not applicable’, therefore the percentage of men is roughly the percentage of women subtracted from 100%.

As can be seen in the table, both groups were very similar, which was expected due to the random distribution of participants.

Table 1: *Demographics of the participants*

Attribute	Value	Treatment	Control	Total
Level of English	Beginner	2%	0%	1%
	Intermediate	16%	10%	14%
	Fluent	47%	42%	45%
	Native Speaker	35%	48%	41%
Education	No schooling completed	0%	3%	1%
	completed primary education	5%	0%	3%
	completed secondary education	26%	29%	27%
	completed post-secondary education	21%	23%	22%
	Bachelor or equivalent	33%	32%	32%
	Master or equivalent	14%	13%	14%
	Doctorate degree or higher	2%	0%	1%
Average age		28.2	29.1	28.9
Percentage of women		46.5%	41.9%	44.6%

6.4 Results

A two-sample test for proportions was conducted for the treatment group (that was talking to the agent), to test the hypothesis that performance after the treatment would be different to the performance before the treatment. The same test was done for the control group. Also, the change in performance was determined. The results can be seen in table 2.

Table 2: *Results of the two-sample tests for proportions*

Tested Category	Condition	Treatment		Control	
		Mean	Std. Err.	Mean	Std. Err.
Sunk Cost Fallacy	before	0.80	0.07	0.63	0.09
	after	1.00	0.00	0.73	0.08
	difference	0.20**	0.07	0.10	0.12
Gambler's Fallacy	before	0.83	0.06	0.70	0.08
	after	0.97	0.03	0.83	0.07
	difference	0.14*	0.07	0.13	0.11
Bayesian Reasoning	before	0.17	0.06	0.07	0.05
	after	0.31	0.08	0.23	0.08
	difference	0.14*	0.07	0.17*	0.09
Belief Bias in Syll. Reas.	before	0.74	0.07	0.57	0.09
	after	0.63	0.08	0.53	0.09
	difference	-0.11	0.11	-0.03	0.13
Regression to the Mean	before	0.34	0.08	0.50	0.09
	after	0.77	0.07	0.63	0.09
	difference	0.43**	0.11	0.13	0.13
Covariation Detection	before	0.71	0.08	0.50	0.09
	after	0.89	0.09	0.53	0.09
	difference	0.17*	0.11	0.03	0.13
Selection Task	before	0.40	0.08	0.37	0.09
	after	0.57	0.08	0.30	0.08
	difference	0.17	0.12	-0.07	0.12

* = $p < 0.05$, ** = $p < 0.01$

Furthermore, a t-test was conducted comparing the results of the treatment group in the second test (after talking to the agent) to the control group in the second text (after reading their texts), to see if there was any significant difference between them. The results are summarized in table 3.

The main findings are presented in the following sections.

Table 3: Results of the t-tests comparing the performance in the second test

Tested Category	Group	Mean	Standard Error
Sunk Cost Fallacy	Treatment	1.00	0.00
	Control	0.73	0.08
	Difference	0.26**	0.08
Gambler's Fallacy	Treatment	0.97	0.03
	Control	0.83	0.07
	Difference	0.14*	0.04
Bayesian Reasoning	Treatment	0.31	0.08
	Control	0.23	0.08
	Difference	0.08	0.05
Belief Bias in Syll. Reasoning	Treatment	0.63	0.08
	Control	0.53	0.09
	Difference	0.10	0.12
Regression to the mean	Treatment	0.77	0.07
	Control	0.63	0.09
	Difference	0.14	0.11
Covariation Detection	Treatment	0.89	0.05
	Control	0.53	0.09
	Difference	0.35**	0.10
Selection Task	Treatment	0.57	0.08
	Control	0.30	0.09
	Difference	0.27*	0.12

* = $p < 0.05$, ** = $p < 0.01$

6.4.1 Sunk Cost Fallacy

In the first test, 7 of the 35 participants (20%) in the treatment group solved the sunk cost fallacy problem incorrectly. After talking to the bot, all 35 solved it correctly. In the control group however, 36% gave a wrong answer in the first test, and still 26% in the second test.

The tests for proportions showed that the participants improved their performance in both groups. The treatment group improved by .2, the control group by .1 units. For the treatment group, the results were very significant ($p = 0.0026$), for the control group however, they were not significant.

The t-test resulted in the treatment group being very significantly better.

6.4.2 Gambler's Fallacy

The treatment group improved from 17% giving the non-normative response to 3%. The control group improved from 30% to 17%.

The tests for proportion showed that both groups improved their performance, the control group only very little less than the treatment group. For the treatment group, the results were significant, for the control group, they were not significant. The t-test showed that the performance of the treatment group after the intervention was significantly better ($p = 0.0282$) than the performance of the control group after their intervention.

6.4.3 Bayesian Reasoning

For the bayesian reasoning, the two probability values closest to the correct one were assigned the '1', and every other value the '0'. 83% of the treatment group and 93% of the control group failed to solve this task correctly in the first test. In the second test, it was 69% and 77%, respectively. The tests for proportions showed that both groups improved, but the improvements were minor. The effect was a little bit stronger for the control group than for the treatment group. For the control group, the effect was significant, for the treatment group, it was not (though almost, $p = 0.0817$). The t-test resulted in the treatment group being not significantly better than the control group.

6.4.4 Belief Bias in Syllogistic Reasoning

In evaluating the belief bias, there was a very small negative effect for the treatment group. The worsened their performance from 26% wrong to 37% wrong, with a difference of -0.114. The effect was not significant. The control group also became worse, although not as much, showing a difference of -.033. This effect was also not significant. This is mainly due to people who were right in the first test, but wrong in the second.

The two-sample t-tests showed no sign of the treatment group being better than the control group as well.

6.4.5 Regression towards the mean

The treatment group improved from 66% errors to 23%, the control group only from 50% to 37%. Two tests for proportion showed a very strong effect for the treatment group (.42) that was highly significant. The effect was positive, too, but less strong (.133) and not significant for the control group.

The two-sample t-test did not indicate the treatment group to be significantly better, however.

6.4.6 Covariation Detection

The treatment group improved from 29% errors to 11%, the control group from 50% to 47%. The tests for proportion showed that there was a small effect (.17) for the treatment group that was also significant, and a very small effect for the control group (.03) that was not significant. The two-sample t-tests resulted in the treatment group being very significantly better.

6.4.7 Wason's Selection Task

For Wason's Selection Task, only those participants who picked the card 'P' and the card 'not-Q' were counted as being correct ('1'), everyone who picked other cards, additional ones, or only one of them was counted as being incorrect ('0'). While the treatment group improved from 60% incorrect solutions to 43%, the control group actually worsened from 63% to 70%. The tests for proportion resulted in a positive effect (.17) that was not significant, but close ($p = 0.07$) for the treatment group, and a close to non-existing negative effect (-0.07) that was also not significant for the control group. The t-test indicated that the treatment group was very significantly better than the control group.

6.4.8 Influence of Age, Gender, Education and Language Proficiency

A linear regression analysis was performed and showed that none of the demographic variables had *any* influence on the improvements gained through working with the agent or taking part in the control group's online course. Not even the level of English accounted for any change in the effectiveness of the interventions.

6.4.9 Cognitive Reflection Test

The participants did not improve in their performance on the cognitive reflection test. The effect resulting from a proportion test comparing their performances in the first test and the second test yielded only a microscopic change that was not significant at all.

6.4.10 Personal Feedback by the Participants

Feedback from participants suggested that they perceived the study as interesting. Some gave feedback like "The chatbot is really nice :)" or "Thank you for this interesting experience, I really liked Liza and her funny and helpful way of teaching. :)". Those answers suggest that the people giving them had a positive social experience with the conversational agent.

A particularly kind person wrote "Ich bin beeindruckt von der umfangreichen, detaillierten Gestaltung von Liza und den Tests und schon gespannt auf die Ergebnisse :)". Also, some participants expressed ideas on how to further improve the agent, for example: "add more actions to the bot, like repeating the question or resetting the answer, so it is possible to change the answer". Those answers suggest that the people giving them were interested and invested thought in the process, which is a good sign. The users did mainly focus on the dialogue structure the agent provided and showed no sign of inappropriately high expectations. They stuck to the tasks they were given by Liza and followed her guidance easily.

6.5 Discussion

6.5.1 Confirmation and Rejection of Hypotheses

Hypothesis 1

This agent was created hoping that the interaction with it would improve the ability of the participants to solve the reasoning tasks. For the sunk cost fallacy, the gambler's fallacy, bayesian reasoning, regression to the mean and covariation detection, there was a significant increase in performance. The two remaining tasks had no significant effect, although the performance on the selection task did noticeably improve. Therefore, the first hypothesis is largely confirmed. The agent was able to improve the reasoning skills of the participants.

Hypothesis 2

The second hypothesis, namely that the non-interactive online course would be able to improve the performance as well, was not fully confirmed. Although there was an increase in performance for nearly every task, only one of them was significant. It can be concluded that the online course was not sufficient to teach reasoning for those tasks.

Hypothesis 3

The work with the agent also proved itself to be more efficient than the control group in many cases. Its effects were stronger and more significant for teaching to avoid the sunk cost fallacy and the gambler's fallacy, the principle of the regression towards the mean and covariation detection. In Wason's selection task, the treatment group improved more than the control group, too, although both were not significant. In conclusion, the interaction with the bot outperformed the non-interactive reading in all those cases. It can be assumed that the agent provided a more efficient intervention, confirming the third hypothesis. Only in bayesian reasoning, the control group improved more than the treatment group, but both effect sizes were very close to each other. In general, teaching bayesian reasoning showed itself to be quite difficult. For the belief bias in syllogistic reasoning, both interventions failed to improve student's performance.

Hypothesis 4

In the sunk cost problem, the gambler's fallacy, the covariation detection task and Wason's Selection tasks, the participants who talked to the agent were not only better than before, but significantly or very significantly better than the control group. The fourth hypothesis can therefore also be accepted with some restrictions. The test subjects were not significantly better in the bayesian reasoning task and the belief bias task, since in both tasks, not much improvement could be gained at all for both groups. They were not significantly better than the control group in the regression to the mean problem (As already stated, the improvement of performance was indeed very significant for this task in the treatment group and not significant for the control group.)

Hypothesis 5

Neither the interaction with the agent nor the reading task helped students improve their general performance on the cognitive reflection test. There is no hint for any of

the interventions to improve general skills in overridden the System 1 response. This hypothesis is rejected, as was expected before. But at least, it helped them overcome certain very typical and common errors, which is already a very satisfying result.

Hypothesis 6

It could not be determined whether the effects would last for two weeks, since only nine participants took part in the third test. All of them performed well on the third, later test, with similar scores to their results on the second test, but this sample is too small to draw any meaningful conclusions. The hypotheses can neither be rejected nor confirmed.

6.5.2 Possible Explanations

Since both tasks with the non-significant results (Bayesian Reasoning and Selection Task) are already considered belonging to the hardest reasoning tasks, having generally very low rates in being solved correctly, and being hard to explain, it can be assumed that a short training was not enough to convey the intended message. After all the subjects spent on average 35-45 minutes with the agent, but this time had to be split between the seven different topics they learned about, leaving only 5-7 minutes on average per topic.

For Wason's Selection Task, it might be necessary to either teach people to view the problem as a permission-based scenario or explain formal logic more thoroughly. For Bayesian reasoning, teaching would probably be more effective if the agent could evaluate single steps of the participants getting to their solution to intervene and correct them if necessary. The other five categories seemed to work well, and in general, one can indeed conclude that the agent improved the user's reasoning skills.

It is very likely that the agent mainly improved the mindware of the students, but not so much the general tendency of using system 2 instead of the intuitive system 1. Therefore, their performance on the CRT did not improve. But a better mindware is still a contributing factor to improved reasoning, of course.

6.6 Limitations

6.6.1 Number and characteristics of participants

The number of 65 acquired participants was higher or similar to some other studies in the field (Okonkwo et al. 2001 (n=12), Shaw et al. 1999a (n=25), Jaques et al. 2009 (n=39)), but more participants would of course offer a more accurate picture of the observed effects. A restriction for the execution of the study was the use of English language by the bot, which required the participants to be able to write and read in English. Instead of relying on students who had time and interest to spend roughly an hour in an experiment at university, the study was conducted online, allowing for participants from different locations to take part in it. It seems highly likely that this allowed for more participants to be acquired.

Of course, no representative sample of any kind could be acquired. Instead, some of

the participants are friends, acquaintances, or acquaintances of acquaintances. A bias towards younger people, especially university students, is therefore inevitable.

6.6.2 Previous Knowledge

Some of the participants already had previous knowledge on the subject, as was assumed in beforehand. Additionally, not many wrong answer were given in the first part of the survey. To give an example: in the experimental condition, only 7 out of 35 participants did answer the sunk cost fallacy in a non-normative way, leaving the agent with only eight opportunities of actually improving somebody's reasoning on this fallacy. All 8 improved.

6.6.3 Scope of Questions

Since time was an important constraint - the whole experiment already took many participants well over 40 minutes, some spent more than an hour on it - it was necessary to keep the tests as short as possible. Therefore, every bias or fallacy (e.g. the sunk cost fallacy) could only be assessed with one question in every test, except for the cognitive reflection test, which was represented by three (short) questions in each test. Errors due to lapses of attention or accidentally ticking the wrong box are therefore possible. It was also not feasible to just compare the gains between the treatment and the control group with respect to every single person and expect any significant, because the difference for e.g. the gambler's fallacy could only have the values -1 (first right, second wrong), 0 (both right) or 1 (first wrong, second right). This would have been much better with e.g. 5 different questions for each topic, but this was not possible due to the already long and exhausting survey of 30-60 minutes. To address a general problem in research on reasoning and decision making, many of the questions are based on hypothetical situations and therefore share all advantages and disadvantages with thought experiments of all kinds. Despite the best efforts to rule out any misinterpretation, a different representation of the situation by the user cannot be excluded.

6.6.4 Time Spent with the Agent

In general, time was the most limiting factor to the experiment. It was crucial to avoid causing the participants to quit due to exhaustion or boredom. Therefore, the agent was restricted to ask only 14 questions (or less, if the user already seems to be very proficient, as was described in section 4.1.4). Liza could therefore test the user's understanding with no more than two questions per topic. As stated before, only 5-7 minutes were spent with each topic. This is simply not much time to actually learn something substantial, but more could not be asked of voluntary test subjects, and paid participants would be more expensive if they would be required to spent more time with the task. Probably, a better effect could be achieved with longer and repeated training with the agent, but those was simply not feasible in the within the study.

Maybe a limitation to an even narrower range of topics would have been helpful, but before actually conducting the study, it was not clear with topics would work best.

6.6.5 Long term effects

Due to the time restrictions for the completion of a bachelor's thesis, it was simply not possible to investigate how much of any learning gains would persist after, say, four months. Also, the only few participants were willing to participate in the follow up test two weeks later. But of course, long term effects are far more useful than temporary improvements.

6.6.6 Minor Concerns

One participant reported a bug that did not allow him to continue with the second part after finishing talking to the agent. Since this error did only occur to one person, it had no major effect. Two users had problems using the website with their smartphones, as already mentioned. Some users did notice occasional typographical or spelling errors. To eradicate those would probably improve the user experience, but did not seem to do any harm for the interaction with the agent.

7 Conclusion

In this work, a conversational agent was presented that was designed to improve skills for a selection of core reasoning tasks. Although it did not have astonishing success in all the tasks, maybe because Wason's Selection Task and Bayesian reasoning are both very complicated concepts to teach, it did show very satisfying results with the other tasks. The agent was clearly more effective than an online course on the same skills, proving its worth as a mentor. Especially having in mind that participants only spent roughly 30 minutes talking to it, it can clearly be viewed as a success.

We can conclude that performance on solving certain reasoning tasks can indeed be improved with the help of pedagogical agents, which is especially fortunate as there are not many teachers in this area, despite its importance for arguably everybody's personal life. Of course, the presented agent is only a prototype developed by a single person and could be improved and extended in many ways, as is described in the following section. Nevertheless, it may be seen as a small hint in the right direction for the application of conversational agent's technologies to education about reasoning and rationality.

The following innovative contributions are presented in this work:

1. A Java architecture for a conversational agent teaching reasoning skills is developed. The agent is able to engage in task-oriented dialogue as well as some chat-oriented conversation and adapts its course on the fly according to the performance of the student.
2. A specialized parser that serves as a tailored solution for parsing user input in this environment is developed.

3. Over sixty different reasoning tasks and their necessary data (complete with explanations, associated small talk, hints, evaluations etc.) are formulated and developed for use in the agent.
4. The general applicability of a conversational agent for teaching the challenging topic of rationality is assessed and evaluated. The effectiveness of teaching for seven different classes of tasks is examined.
5. The advantages of conversational agents in comparison non-interactive teaching methods in the subject of reasoning are evaluated and confirmed.

8 Future Directions

8.1 Improvement of the Agent

8.1.1 Refining the dialogue

Given the nature of the agent, it can always be improved by adding more special parsing options and appropriate responses. For example, if a user would ask the agent about the weather, it would not ‘understand’ it, but, depending on the context, glance over it or ask for a rephrasing fitting to the question asked. By extending the scope of correctly parsed input, the agent can become even more convincing, although there might be diminishing returns for increasingly rare topics.

More elaborate dialogue patterns could be added, too. One user suggested that the agent asks if somebody is actually sure of their answer, allowing for a correction in case of an error. Of course, if the agent asks after every answer the user gives, this would be annoying. A good solution would be one example of how the dialogue could become more elaborate.

8.1.2 Further Personalization of the Learning Plan

Right now, the agent sticks to its course of questions for each user, with the exception that it will allow users to skip questions if they already showed that they understand that topic well. Further improvements could be made by allowing the user to ask the agent for more questions of a certain kind, or the agent suggesting more training in an area where the user seems to lack understanding.

8.1.3 Scaling Up the Learning Experience

The agent could probably be put to a more effective use if it was used not only for a single intervention, but regularly. Storing the past performances of users, remembering their preferences, and specifically targeting the areas where they can improve come to mind as obvious possibilities for expanding the agent.

8.1.4 Expanding the Scope

The agent does only teach students in seven limited and strictly restricted topics. Improvements could include more topics (e.g. addressing some more of the hundreds of known fallacies and biases), offering a more general approach (e.g. giving some philosophical background on knowledge, its representation and a more refined approach on what rationality means), or digging into new areas of the field (e.g. the area of ‘critical thinking’ which includes open-mindedness and discussing the quality of arguments and which was originally intended to be part of the agent, but had to be left out due to time and scale concerns). By those means, the ‘curriculum’ of the agent could be extended in many directions.

8.2 Further Research

First and foremost, it would be very interesting to validate the results with a much larger sample. An expansion of this research could be a study over the course of several weeks or months, to have the users spend much more time with the agent, to see if this approach is more effective. Long term effects could be investigated by reassessing user’s performance after weeks, months or even years. And the suitability of pedagogical agents to help with different biases or concepts, like the one mentioned above, could be tested. A teachable agent instead of a mentor-like one could offer an even more interactive and captivating approach. Also, the use of an animated persona might have an impact on the agent’s effectiveness.

References

- Adcock, A. B., M. H. Duggan, E. Nelson, and C. Nickel (2006). “Teaching effective helping skills at a distance: The development of project CATHIE”. In: *Quarterly Review of Distance Education* 7.4, pp. 349 360.
- Anderson, B. L., S. Williams, and J. Schulkin (2013). “Statistical literacy of obstetrics-gynecology residents”. In: *Journal of graduate medical education* 5.2, pp. 272 275.
- André, E., T. Rist, and J. Müller (1998). “Integrating reactive and scripted behaviors in a life-like presentation agent”. In: *Proceedings of the second international conference on Autonomous agents*, pp. 261 268.
- Arkes, H. R. and C. Blumer (1985). “The psychology of sunk cost”. In: *Organizational behavior and human decision processes* 35.1, pp. 124 140.
- Banchs, R. E. and H. Li (2012). “IRIS: a chat-oriented dialogue system based on the vector space model”. In: *Proceedings of the ACL 2012 System Demonstrations*, pp. 37 42.
- Bates, J. et al. (1994). “The role of emotion in believable agents”. In: *Communications of the ACM* 37.7, pp. 122 125.
- Baylor, A. (2000). “Beyond butlers: Intelligent agents as mentors”. In: *Journal of Educational Computing Research* 22.4, pp. 373 382.

- Baylor, A. and Y. Kim (2005). "Simulating instructional roles through pedagogical agents". In: *International Journal of Artificial Intelligence in Education* 15.2, pp. 95 115.
- Beach, L. R. and R. G. Swensson (1967). "Instructions about randomness and run dependency in two-choice learning". In: *Journal of Experimental Psychology* 75.2, p. 279.
- Berwick, D. M., H. V. Fineberg, and M. C. Weinstein (1981). "When doctors meet numbers". In: *The American journal of medicine* 71.6, pp. 991 998.
- Bickmore, T. W. and R. W. Picard (2003). "Subtle Expressivity by Relational Agents". In: *In Proceedings of the CHI 2003 Workshop on Subtle Expressivity for Characters and Robots*.
- Bickmore, T., D. Schulman, and L. Yin (2009). "Engagement vs. deceit: Virtual humans with human autobiographies". In: *International Workshop on Intelligent Virtual Agents*, pp. 6 19.
- Blair, K., D. Schwartz, G. Biswas, and K. Leelawong (2007). "Pedagogical agents for learning by teaching: Teachable agents". In: *Educational Technology* 47.1, p. 56.
- Bruine de Bruin, W., A. M. Parker, and B. Fischhoff (2007). "Individual differences in adult decision-making competence". In: *Journal of personality and social psychology* 92.5, p. 938.
- Caplan, D. (2006). *Management accounting concepts and techniques*. URL: <http://denniscaplan.fatcow.com/TOC.htm> (visited on 08/02/2016).
- Casscells, W., A. Schoenberger, and T. B. Graboys (1978). "Interpretation by physicians of clinical laboratory results". In: *New England Journal of Medicine* 299.18, pp. 999 1001.
- Chase, C. C., D. B. Chin, M. A. Opezzo, and D. L. Schwartz (2009). "Teachable agents and the protégé effect: Increasing the effort towards learning". In: *Journal of Science Education and Technology* 18.4, pp. 334 352.
- Cheng, P. W., K. J. Holyoak, R. E. Nisbett, and L. M. Oliver (1986). "Pragmatic versus Syntactic Approaches to Training Deductive Reasoning." In: *Cognitive psychology* 18, pp. 293 328.
- Chiesi, F., C. Primi, and K. Morsanyi (2011). "Developmental changes in probabilistic reasoning: The role of cognitive capacity, instructions, thinking styles, and relevant knowledge". In: *Thinking & Reasoning* 17.3, pp. 315 350.
- Choi, S. and R. E. Clark (2006). "Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language". In: *Journal of Educational Computing Research* 34.4, pp. 441 466.
- Clark, H. H. (1999). "How do real people communicate with virtual partners". In: *Proceedings of 1999 AAAI Fall Symposium, Psychological Models of Communication in Collaborative Systems*, pp. 43 47.
- DeVault, D., R. Artstein, G. Benn, T. Dey, et al. (2014). "SimSensei Kiosk: A virtual human interviewer for healthcare decision support". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1061 1068.

- DeVault, D., J. Mell, and J. Gratch (2015). "Toward natural turn-taking in a virtual human negotiation agent". In: *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*. AAAI Press, Stanford, CA.
- Dirkin, K. H., P. Mishra, and E. Altermatt (2005). "All or nothing: Levels of sociability of a pedagogical software agent and its impact on student perceptions and learning". In: *Journal of Educational Multimedia and Hypermedia* 14.2, p. 113.
- Driscoll, M. P. and M. P. Driscoll (2005). "Psychology of learning for instruction". In: Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: Problems and opportunities". In:
- Epstein, D. and E. Reategui (2014). "A pedagogical agent with embedded data mining functions to support collaborative writing". In: *learning* 12.13, p. 14.
- Evans, J. S. B., J. L. Barston, and P. Pollard (1983). "On the conflict between logic and belief in syllogistic reasoning". In: *Memory & Cognition* 11.3, pp. 295 306.
- Evans, J. S. B. (1984). "Heuristic and analytic processes in reasoning". In: *British Journal of Psychology* 75.4, pp. 451 468.
- Evans, J. S. B. and D. E. Over (1996). *Rationality and Reasoning*. Psychology Press.
- Evans, J. S. B. and K. E. Stanovich (2013). "Dual-process theories of higher cognition advancing the debate". In: *Perspectives on psychological science* 8.3, pp. 223 241.
- Fogg, B. J. and C. Nass (1997). "Silicon sycophants: the effects of computers that flatter". In: *International Journal of Human-Computer Studies* 46.5, pp. 551 561.
- Fong, G. T., D. H. Krantz, and R. E. Nisbett (1986). "The effects of statistical training on thinking about everyday problems". In: *Cognitive psychology* 18.3, pp. 253 292.
- Fonte, F. A., J. Carlos, B. Rial, and M. L. Nistal (2009). "TQ-Bot: an AIML-based tutor and evaluator bot". In: *Journal of Universal Computer Science* 15.7, pp. 1486 1495.
- Forsyth, D. R. (2008). "Self-serving bias". In: *International Encyclopedia of the Social Sciences*.
- Frederick, S. (2005). "Cognitive reflection and decision making". In: *The Journal of Economic Perspectives* 19.4, pp. 25 42.
- FreeLing Home Page (2016). URL: <http://nlp.lsi.upc.edu/freeling/node/1> (visited on 08/20/2016).
- Friedman, D., K. Pommerenke, R. Lukose, G. Milam, et al. (2007). "Searching for the sunk cost fallacy". In: *Experimental Economics* 10.1, pp. 79 104.
- Graesser, A. C., P. Chipman, B. C. Haynes, and A. Olney (2005). "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue". In: *Education, IEEE Transactions on* 48.4, pp. 612 618.
- Graesser, A. C., M. Jeon, and D. Dufty (2008). "Agent technologies designed to facilitate interactive knowledge construction". In: *Discourse processes* 45.4-5, pp. 298 322.
- Griffin, D. and A. Tversky (1992). "The weighing of evidence and the determinants of confidence". In: *Cognitive psychology* 24.3, pp. 411 435.
- Gross, S., B. Mokbel, B. Hammer, and N. Pinkwart (2015). "Learning feedback in intelligent tutoring systems". In: *KI-Künstliche Intelligenz* 29.4, pp. 413 418.
- Gulz, A. (2005). "Social enrichment by virtual characters differential benefits". In: *Journal of Computer Assisted Learning* 21.6, pp. 405 418.

- Gulz, A., M. Haake, A. Silvervarg, B. Sjöden, et al. (2011). “Building a Social Conversational Pedagogical Agent: Design Challenges and”. In: *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices*, p. 128.
- Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge University Press.
- Haferkamp, N., N. C. Kraemer, C. Linehan, and M. Schembri (2011). “Training disaster communication by means of serious games in virtual environments”. In: *Entertainment Computing 2.2*, pp. 81–88.
- Halpern, D. F. (2002). *Thought and knowledge: An introduction to critical thinking*. Routledge.
- Hitchcock, D. (2003). “The effectiveness of computer-assisted instruction in critical thinking”. In: *Informal Logic: Reasoning and Argumentation in Theory and Practice* Vol. 24, No. 3, pp. 183–217.
- Hoffrage, U. and G. Gigerenzer (1998). “Using natural frequencies to improve diagnostic inferences”. In: *Academic medicine* 73.5, pp. 538–540.
- Hooshyar, D., R. B. Ahmad, M. Yousefi, F. D. Yusop, et al. (2015). “A flowchart-based intelligent tutoring system for improving problem-solving skills of novice programmers”. In: *Journal of Computer Assisted Learning* 31.4, pp. 345–361.
- Inglis, M. and A. Simpson (2004). “Mathematicians and the selection task”. In: *Proceedings of the 28th International Conference on the Psychology of Mathematics Education*. Vol. 3, pp. 89–96.
- Jaques, P. A., S. Pesty, and R. Vicari (2003). “An animated pedagogical agent that interacts affectively with the student”. In: *AIED 2003, shaping the future of learning through intelligent technologies*, pp. 428–430.
- Jaques, P. A., M. Lehmann, and S. Pesty (2009). “Evaluating the affective tactics of an emotional pedagogical agent”. In: *Proceedings of the 2009 ACM symposium on Applied Computing*, pp. 104–109.
- Jungermann, H. (1976). *Rationale Entscheidungen*. Vol. 10. Psychologisches Kolloquium. Verlag Hans Huber Bern Stuttgart Wien. ISBN: 3-456-80267-6.
- Kahneman, D. (2003). “A perspective on judgment and choice: mapping bounded rationality”. In: *American psychologist* 58.9.
- Kahneman, D. and A. Tversky (1996). “On the reality of cognitive illusions”. In:
- Kim, Y. and A. Baylor (2006). “A social-cognitive framework for pedagogical agents as learning companions”. In: *Educational Technology Research and Development* 54.6, pp. 569–596.
- Kim, Y. and A. Baylor (2007). “Pedagogical agents as social models to influence learner attitudes”. In: *Educational Technology* 47.1, pp. 23–28.
- Klaczynski, P. A. (2001). “Analytic and Heuristic Processing Influences on Adolescent Reasoning and Decision-Making”. In: *Child development* 72.3, pp. 844–861.
- Kramer, M., R. Yaghoubzadeh, S. Kopp, and K. Pitsch (2013). “A conversational virtual human as autonomous assistant for elderly and cognitively impaired users? Social acceptability and design considerations”. In: *GI-Jahrestagung*, pp. 1105–1119.

- Kumar, A. N. (2016). “Providing the Option to Skip Feedback in a Worked Example Tutor”. In: *International Conference on Intelligent Tutoring Systems*, pp. 101–110.
- Kumar, R., G. Gweon, M. Joshi, Y. Cui, et al. (2007). “Supporting students working together on math with social dialogue”. In: *SLaTE*, pp. 96–99.
- Lane, D., J. Lu, C. Perese, and E. Zitek (2016). *Online Statistics: An Interactive Multimedia Course of Study: An introductory-level statistics book*. URL: http://onlinestatbook.com/version_1.html (visited on 08/03/2016).
- Lane, H. C., C. Cahill, S. Foutz, D. Auerbach, et al. (2013). “The effects of a pedagogical agent for informal science education on learner behaviors and self-efficacy”. In: *International Conference on Artificial Intelligence in Education*, pp. 309–318.
- Larrick, R. P. (2004). “Debiasing”. In: *Blackwell handbook of judgment and decision making*, pp. 316–337.
- Larrick, R. P., J. N. Morgan, and R. E. Nisbett (1990). “Teaching the use of cost-benefit reasoning in everyday life”. In: *Psychological Science* 1.6, pp. 362–370.
- Latham, A., K. Crockett, D. McLean, and B. Edmonds (2012). “A conversational intelligent tutoring system to automatically predict learning styles”. In: *Computers & Education* 59.1, pp. 95–109.
- Lester, J. C., S. A. Converse, S. E. Kahler, S. T. Barlow, et al. (1997). “The persona effect: affective impact of animated pedagogical agents”. In: *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pp. 359–366.
- Markovits, H. and G. Nantel (1989). “The belief-bias effect in the production and evaluation of logical conclusions”. In: *Memory & Cognition* 17.1, pp. 11–17.
- Mikic, F. A., J. C. Burguillo, M. Llamas, E. Rodriguez, et al. (2009). “CHARLIE: An AIML-based Chatterbot which Works as an Interface among INES and Humans”. In: *EAAEEIE Annual Conference, 2009*, pp. 1–6.
- More, D. J. (2012). “Development and Evaluation of Web-Based Animated Pedagogical Agents for Facilitating Critical Thinking in Nursing”. In: *Nursing education perspectives* 33.2, pp. 116–120.
- Moreno, R., R. E. Mayer, H. A. Spires, and J. C. Lester (2001). “The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?” In: *Cognition and instruction* 19.2, pp. 177–213.
- Morewedge, C. K., H. Yoon, I. Scopelliti, C. W. Symborski, et al. (2015). “Debiasing Decisions Improved Decision Making With a Single Training Intervention”. In: *Policy Insights from the Behavioral and Brain Sciences*, p. 2372732215600886.
- Nass, C., Y. Moon, and P. Carney (1999). “Are People Polite to Computers? Responses to Computer-Based Interviewing Systems¹”. In: *Journal of Applied Social Psychology* 29.5, pp. 1093–1109.
- Nass, C., Y. Moon, B. J. Fogg, B. Reeves, et al. (1995). “Can computer personalities be human personalities?” In: *International Journal of Human-Computer Studies* 43.2, pp. 223–239.
- Nass, C., J. Steuer, and E. R. Tauber (1994). “Computers are social actors”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 72–78.

- Nio, L., S. Sakti, G. Neubig, T. Toda, et al. (2014). “Developing non-goal dialog system based on examples of drama television”. In: *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, pp. 355 361.
- Nisbett, R. E., G. T. Fong, D. R. Lehman, and P. W. Cheng (1987). “Teaching reasoning”. In: *science* 238.4827, pp. 625 631.
- Nisbett, R. E., D. H. Krantz, C. Jepson, and Z. Kunda (1983). “The use of statistical heuristics in everyday inductive reasoning”. In: *Psychological Review* 90.4, p. 339.
- Okonkwo, C. and J. Vassileva (2001). “Affective pedagogical agents and user persuasion”. In: *HCI*, pp. 397 401.
- Ortony, A. (2002). “On making believable emotional agents believable”. In: *Trappl et al.(Eds.)(2002)*, pp. 189 211.
- Perkins, D. N. (1989). *Mindware: The new science of learnable intelligence*. Association for Supervision and Curriculum Development.
- Picard, R. W. (2003a). “Affective computing: challenges”. In: *International Journal of Human-Computer Studies* 59.1, pp. 55 64.
- Picard, R. W. (2003b). “What does it mean for a computer to “have” emotions”. In: *Emotions in humans and artifacts*, pp. 213 235.
- Pinkwart, N. and B. M. McLaren (2012). *Educational technologies for teaching argumentation skills*. Bentham Science Publishers.
- Reeves, B. and C. Nass (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press Cambridge, UK.
- Roney, C. J. and L. M. Trick (2003). “Grouping and gambling: a Gestalt approach to understanding the gambler’s fallacy”. In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 57.2, p. 69.
- Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Psychology Press.
- Sedlmeier, P. and G. Gigerenzer (2001). “Teaching Bayesian reasoning in less than two hours”. In: *Journal of experimental psychology: general* 130.3, p. 380.
- Shaw, E., R. Ganeshan, W. L. Johnson, and D. Millar (1999a). “Building a case for agent-assisted learning as a catalyst for curriculum reform in medical education”. In: *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 509 516.
- Shaw, E., W. L. Johnson, and R. Ganeshan (1999b). “Pedagogical agents on the web”. In: *Proceedings of the third annual conference on Autonomous Agents*, pp. 283 290.
- Shen, E. (2016). “Effects of agent emotional support and cognitive motivational messages on math anxiety, learning, and motivation”. In:
- Stanovich, K. E. (2009a). “Rational and irrational thought: The thinking that IQ tests miss”. In: *Scientific American Mind* 20.6, pp. 34 39.
- Stanovich, K. E. (2009b). *What intelligence tests miss: The psychology of rational thought*. Yale University Press.
- Stanovich, K. E. (2012). “On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning”. In: *The Oxford handbook of thinking and reasoning*, pp. 343 365.

- Stanovich, K. E. and R. F. West (1998). "Individual differences in rational thought". In: *Journal of experimental psychology: general* 127.2, p. 161.
- Stanovich, K. E. and R. F. West (1999). "Individual differences in reasoning and the heuristics and biases debate". In:
- Stanovich, K. E. and R. F. West (2000). "Individual differences in reasoning: Implications for the rationality debate?" In: *Behavioral and brain sciences* 23.5, pp. 645 726.
- Sweller, J. (1994). "Cognitive load theory, learning difficulty, and instructional design". In: *Learning and instruction* 4.4, pp. 295 312.
- Sweller, J. (2004). "Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture". In: *Instructional science* 32.1-2, pp. 9 31.
- Thaler, R. (1980). "Toward a positive theory of consumer choice". In: *Journal of Economic Behavior & Organization* 1.1, pp. 39 60.
- Toplak, M. E., R. F. West, and K. E. Stanovich (2011). "The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks". In: *Memory & Cognition* 39.7, pp. 1275 1289.
- Toplak, M. E., R. F. West, and K. E. Stanovich (2014). "Assessing miserly information processing: An expansion of the Cognitive Reflection Test". In: *Thinking & Reasoning* 20.2, pp. 147 168.
- Tunstall, S. L. (2016). "Online Statistics: An Interactive Multimedia Course of Study". In: *Numeracy: Advancing Education in Quantitative Literacy* 9.1.
- Tversky, A. and D. Kahneman (1974). "Judgment under uncertainty: Heuristics and biases". In: *science* 185.4157, pp. 1124 1131.
- van Gelder, T. (2001). "How to improve critical thinking using educational technology". In: *Meeting at the crossroads: Proceedings of the 18th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*, pp. 539 548.
- van Gelder, T. J. (2000). "Learning to reason: A Reason!-Able approach". In: *Cognitive Science in Australia, 2000: Proceedings of the Fifth Australasian Cognitive Science Society Conference. Adelaide: Causal*.
- van Mulken, S., E. André, and J. Müller (1998). "The persona effect: how substantial is it?" In: *People and computers XIII*. Springer, pp. 53 66.
- Veletsianos, G. and G. S. Russell (2014). "Pedagogical agents". In: *Handbook of research on educational communications and technology*. Springer, pp. 759 769.
- Walton, D. (2000). "Problems and useful techniques: My experiences in teaching courses in argumentation, informal logic and critical thinking". In: *Informal Logic* 20.2.
- Wason, P. C. (1968). "Reasoning about a rule". In: *The Quarterly journal of experimental psychology* 20.3, pp. 273 281.
- Weizenbaum, J. (1966). "ELIZA a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1, pp. 36 45.

Woo, H. L. (2009). "Designing multimedia learning environments using animated pedagogical agents: factors and issues". In: *Journal of Computer Assisted Learning* 25.3, pp. 203-218.

Appendices

Asking a Question

Data: int time: how long the agent will wait for answers

Result: a String containing the user's answer

```
timer = 0;
ui.listen_timeout(time);
// makes UI wait for time seconds for an input
while user did not answer do
    increment timer;
    wait for if timer < 5 then
        | Utter short gap-filling comment;
    end
    if timer >= 5 && timer < 8 then
        | Make a worried statement about the user being absent;
    end
    if timer == 12 then
        | Ask the user if they want to continue with the task at all;
        if askbreak() then
            | // askbreak() negotiates with the user about ending the
            | // programm. Return true if the user wants to continue.
            | timer = 0;
            | tell user the original question again;
        else
            | Tell the user goodbye message;
            | Shut down;
        end
    end
end
end
parse the answer;
if user asked the agent to wait then
    | return askForever(60);; // recursively call askForever, this time
    | waiting 1 minute
end
if user answered ambiguous then
    | /* when the input could be an answer to the gap-filling or
    |    worried statements, or to the original question */
    | Express being happy the user is still there;
    | Ask to what the answer referred to;
    if answer referred to original question then
        | return answer;
    end
    | Ask question again; return askForever(time);
end
return answer;
```

Algorithm 1: Getting an answer to a question from the user

A Sample Question

type: 2 ; // gambler's fallacy
text: "William and Susan are buying lottery tickets and choose their numbers. William chooses 6, 14, 22, 35, 38 and 40. Susan chooses 1, 2, 3, 4, 5 and 6. Regardless of how much money they will win, what do you think?";
question: "Whose numbers are more likely to be the winning numbers?";
clarification: "Who is more likely to win, William or Susan?";
context: lottery;
positive tags: both,equal,same;
negative tags: difference,william,susan,more random;
introduction: "Do you like gambling?";
introduction answer 1: "Well, may the odds be always in your favour!";
introduction answer 2: "I see. I, personally, am always calculating the odds and estimated returns. But it's easier if you're a computer, of course.";
answer correct: "That's exactly correct. You have a clear mind when it comes to gambling.";
answer incorrect: "Sorry, this was not correct.";
explanation: "Random chance is just as likely to produce 1 as a first number as it is to produce 6, and the same goes for the second number (where every number but the first can be picked with equal probability), and so on. The single numbers are in no way connected, they are independent from each other. And every possible pattern of statistically independent events like in the lottery is as likely as every other.";
hint: "Think about this: does the 3 being drawn have anything to do with whether or not the 2 was drawn? Do they have some kind of connection, or are they in fact independent of each other?";

Algorithm 2: An exemplary question

If the agent already knows that the user likes gambling, it will not ask the introductory question, but instead get one of the following intros from the phrasebase:

- "As I know you're a fan of gambling, you will probably like this!"
- "You said you liked gambling, didn't you? This could be interesting for you."
- "As someone who's into gambling, you could like this."
- "Since you said you enjoy to gamble with probabilities, this could be interesting to you."
- "Since you said you like gambling, I thought about this..."

If the user already said he is not very much into gambling, the agent will acquire one of the following phrases from the phrasebase instead:

- "I know you're not really a gambling person, but I'd like to have your opinion on this."
- "Since you are critical of gambling, it would be interesting to see what your opinion on this one is."

- “I know you are sceptical of gambling. Therefore I would like to hear your opinion of a certain situation.”
- “It’s another question about gambling, sorry. But I have to find this out!”
- “Well, I know you’re not that much into gambling, but I have to ask you another question about it. Sorry!”

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, den 9. September 2016

.....