Teaching Without Learning?
Exploring the Protégée-effect and Perceived Intelligence in a Teachable Agent Software

Undervisning utan lärande?
Ett utforskande av protégée-effekten och den uppfattade intelligensen hos en artificiell lärande agent

Kristian Månsson

handledare / supervisors

Agneta Gulz
Magnus Haake

KOGM20

2017-06-09
Teaching Without Learning? Exploring the Protégée-effect and Perceived Intelligence in a Teachable Agent Software

Kristian Månsson
civ04kma@student.lu.se

Studies have shown that learning by teaching using teachable virtual agents in educational software benefits learning through a greater sense of responsibility and metacognitive training. A theoretical concept aimed at describing these beneficial factors is the protégée-effect i.e. that the student makes greater effort towards their learning by a feeling of responsibility towards their agent, adopting an incrementalist view of knowledge, and acquiring an ego-protective buffer by having the possibility to blame the agent rather than oneself for failures.

The aim of this study is to investigate whether two different AI settings for a teachable agent in a teachable agent software affects the student’s perception of the agent and the student’s learning.

94 Swedish students in grades 5 and 6 in a primary school, randomly assigned one of two groups of 47 students each used the Guardians of History teachable agent software with two different agent learning behaviours on two occasions. The perceived intelligence of the agent, the protégée-effect and the student’s knowledge gained were measured using a questionnaire. No significant difference between the groups were found regarding perceived agent intelligence or knowledge attained, which hint towards a design choice where the narrative and not the learning model of the teachable agents is in focus. That is, when investing or designing a teachable agent, the artificial intelligence that instantiates the learning model of the teachable agent seems to be of lesser importance. However, a strong correlation ($r = 0.64, p < 0.01$) between the perceived agent intelligence and the protégée-effect was present, indicating that these either influence each other strongly or are measurements of the same underlying phenomena. A correlation designers and researchers working with teachable agents should consider. A lack of correlation between PE and score in the knowledge test indicates that the knowledge test is too easy or that the validity of the measurements is questionable. More research is needed to investigate the validity of the measurement techniques.

Keywords: Learning By Teaching, Peer Tutoring, Teachable Agents, Perception of Intelligence, Virtual Agents, Pedagogical Agents

1 Introduction

The ascription or perception of intelligence in others is one of the most fundamental human experiences. Philosophers such as Daniel Dennett discuss the “intentional stance” whilst phenomenologists like Martin Heidegger writes about Mitsein (“being-with”) to name just a couple of influential schools of thought on this subject. Alan Turing even operationalised the question into an empirical one, the now famous so-called ”Turing test”. AI-researchers and cognitive scientists alike have been investigating how this behaviour manifests itself in a setting where ”the other” is artificial and the usage of artificial agents in learning environments gives an interesting case study of this very general - and fundamental - subject.

Artificial agents that are teachable (henceforth Teachable Agent, or TA) have been shown to be beneficial in terms of learning as well as promote metacognitive training. The different TAs presented in this paper are however differently implemented in important aspects, and as the causation of the beneficial learning functions and some of the theoretical underpinnings of the field is underdeveloped, more research is needed to provide both theoretical development as well as good design guidelines. The goal of this study is thus to contribute to the field by investigating the relationship between how the agent is implemented, perceived and any beneficial learning outcomes. The research questions under investigation are more specifically if a TA with associative learning algorithm rather than a recency learning algorithm give more perceptive effects regarding intelligence of the TA, if a greater perceived agent intelligence correlate with a stronger protégée-effect and if the AI of the teachable agent correlates with the learning outcome.

This paper will present an overview of the field of teachable agents and the software environments in which they are provided - so-called Teachable Agent Learning Environments (henceforth TALE) - with examples of how researchers have tackled the design of TAs and TALEs, present relevant research and give an example of how this can be studied with regard to how the TA is implemented. The study focuses on how the relation between how the TA behaves and how it is perceived and how the beneficial effects of utilising a TALE relates to this. This is done by a questionnaire and a knowledge test on 5th and 6th-graders working with two different TA-implementations in a TALE. The paper is concluded with suggestions of further research on the subject.

Learning by Teaching and Peer Tutoring

Virtual teachable agents functions within the context of learning by teaching, which has a long history. As early as AD 65, Seneca the Younger wrote to his friend Lucilius “hominis dum docent discunt” or “humans learn while
they teach” (Seneca the younger, see Gummere, Seneca & Corcoran, 2006, 7.9), and has also been discussed by influential thinkers such as Vygotsky and Bruner (see Wood, Bruner & Ross, 1976). Learning by teaching increases the motivation towards learning and trains cognitive abilities such as reflection on gained knowledge (Ogan, Finkelstein, Mayfield, D’Adamo, Matsuda, & Cassell, 2012) as well as helps the student to reach a deeper understanding of the topic due to their feeling of responsibility as a teacher to both gain a clear understanding of the learning material as well as conceptualizing the learning content (Biswas, Katzberger & Bransford, 2001). Tutors gain a better understanding of the subjects and a more positive attitude towards them (Cohen, Kulik & Kulik, 1982), and in the process gain school morale and confidence, especially low performers (Frager & Stern, 1970). Bargh & Schul (1980) showed in an experiment on 42 undergraduates that those students who studied expecting to teach other students in preparation for a test performed significantly better than students who studied for themselves.

Peer tutoring refers to a situation where the role of the teacher is assigned and the teacher is untrained and where the differences in knowledge of the teacher and the tutee can be small (Pareto, 2014). The tutor gains knowledge through the preparation of the teaching session through revision of her own understanding of the domain knowledge, through self-monitoring of her comprehension, integration of new and prior knowledge, and elaboration and construction of knowledge (Roscoe & Chi, 2007). Pareto (2014) also points out that another powerful aspect of peer-tutoring is that the tutee asks questions that forces the tutors to take stand on their knowledge or disclose lacks in the domain knowledge of the tutee. Indeed, as Ur & VanLehn (1995) notes, the basic interaction between a tutor and a tutee is feedback, hints, and asking questions. Interaction between an tutor and a tutee should be observed and discussed with these elemental interaction types in mind.

Teachable (Virtual) Agents

Teachable virtual agents may provide the solution to some of the potential problems of utilising learning by teaching through peer tutoring in a school setting (Kirkegaard, 2016). It could be argued that that no real student comes to harm through a failed teaching session (Kirkegaard, Gulz, & Silvervarg, 2014), i.e. that student may fall behind or learn erroneous facts due to failed tutoring. Kirkegaard (2016) also points out that finding matching peers could be difficult. A TALE can be adapted to match the level of each student with regards to domain knowledge and metacognitive competence. The usage of TA in a school setting has the additional benefit as well as provide researches with the opportunity to study peer learning and peer tutoring behaviour in a controlled manner (Ogan, Finkelstein, Mayfield, D’Adamo, Matsuda, & Cassell, 2012).

According to Pareto, Schwartz & Svensson (2009) there is a distinction between a learning companion that can learn and an agent that can be taught. An important distinction to have in mind as they are not contingent upon each other. A teachable agent only has to appear to be taught. However, as Chan & Chou (1997) notes, there are two main objectives to a TALE: to model a progression in domain knowledge of the tutee and to make that progression understandable for the human teacher in order to monitor the teaching activities. In their work they have opted for an emulation of an average student learner rather than machine learning algorithms.

Blair, Schwartz, Biswas & Leelawong (2006) states that there are four criteria that a TALE should demonstrate in order to effectively maximize the positive learning effects of a TA.

1. An explicit and well structured visual representation of the agent’s learning that allows for the concrete externalization of the student’s learning. Externalization is what gives the relieving effect of the metacognitive load and also aims at showing the link between good learning strategies and results.

2. A central aspect of a TA is that the agent must be able to do independent reasoning or in some other way produce statements based on her knowledge, as this is the point of having an agent that is teachable. This can give a self-regulatory effect if the agent spontaneously comes to conclusions that seems unreasonable, something that it can point out to the student.

3. The agent must also in some sense model productive learning behaviours because the point of any learning software is to guide the user to a more effective learning behaviour. Thus, one of the main goals of a TALE should either be to directly through, for example, an implementation of one or more learning models, or indirectly, through a narrative provided in a TALE, demonstrate good learning strategies. A TALE should also provide some guidance and mentoring about how the student should best go for teaching his agent.

4. To provide an environment that supports learning providing the necessary tools for the learning process. It is also important that a TALE provides goals that are important for the learner to achieve, thus giving rise to different motivational factors. It can mean simple goals such as winning a quiz, collecting points or making progression in a larger narrative.

Studies have also shown that a TA helps children use and develop their metacognitive skills (Schwartz et al., 2009, Biswas et. al, 2005, Kinnebrew, Biswas, Sulzer & Taylor, 2010). Metacognitive skills have proven to be of great importance to learning. While some of these skills can be developed along with other cognitive abilities as children grow without any training. Others need to be trained, especially self-monitoring and self-regulatory behaviours, both of which are essential for good learning behaviours (Schneider, 2008). Self-regulation is the ability to monitor their own thoughts and ensure that the thought processes deliver the desired results. Winnie & Nesbit (2009) believes that students generally are very untrained and generally quite poor in taking advantage of their metacognitive skills. Students in general are bad at deciding what they actually know, available effective study techniques and how to derive strategies from situations where they are making mistakes. The potential gain to be had from training these abilities can be substantial.
A TALE might help to solve two psychological challenges in metacognitive training. One of the is the heightened cognitive load on the student. When the student is tasked with both monitoring her own thinking and learning behaviour and at the same time partaking in a learning activity, the dual-task cognitive strain might be too great for any metacognitive training to take place. A TA externalises the student’s cognition by being responsible for any problem solving and through that alleviates the cognitive load on the student, and frees up resources for metacognition. Another challenge is a motivational one. Metacognitive reflection is for the student additional work beyond what is typically required for completing learning activities. Students often skim readings and fail the check on their own understanding. A teacher however is responsible for the students understanding an performance which provides motivation for a student using a TALE to engage in metacognitive reasoning (Schwartz et al., 2009).

One of the earliest examples of a TALE is DENISE (Development Environment for an Intelligent Student in Economics), which is developed for for casual reasoning in economics (Nichols, 1994). DENISE is based on the reasoning behind Palthepu, Greer & McCallas’ (1991) suggestion for a TALE where dialogue-driven construction of an initially empty semantic network is gradually constructed as the knowledge model of the system, where several dialogue heuristics (such as find contradictions, check completeness and confirm existing knowledge) are utilised to guide the tutor.

One of the earliest examples of a TALE is DENISE (Development Environment for an Intelligent Student in Economics), which is developed for for casual reasoning in economics (Nichols, 1994). DENISE is based on the reasoning behind Palthepu, Greer & McCallas’ (1991) suggestion for a TALE where dialogue-driven construction of an initially empty semantic network is gradually constructed as the knowledge model of the system, where several dialogue heuristics (such as find contradictions, check completeness and confirm existing knowledge) are utilised to guide the tutor.

According to Nichols (1994), a true tabula rasa-system such as DENISE left the tutors a feeling of discomfort as they where experiencing a feeling of just “feeding the system knowledge”. Hence another approach was adopted by Ur & VanLehn (1995) when designing STEPs. The system for learning how to solve qualitative problems in physics has a complete domain knowledge model and the learning behaviour of the TA is scripted. The tutor gives feedback to the TA by answering simple questions and showing correct steps in the solution of the physics problems (see Figure 2).

The same kind of scripted behaviour can be seen in the reciprocal tutoring system for computer programming in Lisp that was developed by Chan & Chou (1997). The software had the capability for a human and the software to take turns with tutoring and being tutored on alternating programming problems. The TA in this software is described as having the “psychological plausibility” of a learner for the human tutor, but they did not implement any machine learning algorithms. The TA models the learning behaviour with gradually diminishing range errors in the problem solutions suggested.

A well-studied TALE is Betty’s Brain (BB), which is an implementation primarily used in the STE AM subjects and is developed by The Teachable Agents Group at Vanderbilt University (Leeawong, Biswas & Isis, 2008). Their approach is very similar to that of DENISE, in that it couples knowledge concepts in causal chains, but one important difference in that it provides an explicit view of the learning and knowledge model of the TA (see Figure 3).

In BB, the student is tasked with “building Betty’s brain”. The brain consists of a conceptual map with links between the concepts. The concepts consist of nouns such as organic waste, oxygen, and bacteria, representing relevant learning material. The links explain how these concepts in the map are connected in a causal context. These are verbs such as “broken down by,” “consuming” and “increases”. The couplings are also given a positive or negative weight, which indicates whether the verbs are positive or negative, which means “increases” or “decreases.” The concept map is then used to draw conclusions about how different concepts are linked by Betty following the pathways that are formed. An example of a chain is: ammonia is metabolised by (+) bacteria - consuming (-) - acid.

Building Betty’s brain thus means, in some sense, building a machine that can perform certain tasks. For help, the student has a mentor character (Mr. Davis) that is helping the student to teach Betty the best way possible. There is no strong narrative; The student is expected to train Betty until the student is sure that Betty can answer a quiz or compete in a quiz show where three other agents trained by other students face each other.

In terms of externalization, it’s hard to make it more explicit than in the case of BB, since the concept map is exactly the entire agent’s thinking. Betty’s thinking thus becomes completely deterministic in the sense that the agent can never make mistakes. The conclusions drawn as such may be wrong, but there is no uncertainty in the process. Betty never experiences any “aha”-experiences or is gradually becoming more capable of solving certain tasks.

Kinnebrew, Biswas, Sulcer & Taylor (2010) used
Betty’s Brain to assess the metacognitive training when utilising a TALE. They showed an increase in use of metacognitive monitoring behaviors and also observed more advanced and focused monitoring behaviors who where given strategy feedback from a mentor agent.

Another approach regarding the externalization of the learning of the TA is provided in Squares Family (SF) which is a software for training basic arithmetic understanding (Pareto, Schwartz & Svensson, 2009). SF is a type of card game in which the student not only teaches the logic of arithmetic, but also game rules and effective playing strategies as well as the graphic model of the game. The player teaches the TA by example and the TA gives multiple choice questions to the player about both playing strategies as well as the domain knowledge for direct corrective feedback see Figure 4. The agent’s gained knowledge is presented to the player as a barometer for each of the skills attained.

Pareto, Schwartz & Svensson (2009) showed using Squares Family that low performing students gained significant advantages when using a TA compared to not. Pareto (2014), again using Squares Family, enrolled 443 students to show that using a TALE compared to regular instruction in an authentic classroom setting resulted in significant learning gains. A similar study conducted on Betty’s Brain resulted in a similar result; 58 elementary school children using a TA exhibited “added-value” learning that did not adversely affect the “basic-value” they gained from their regular lessons (Chin, Dohmen, Cheng, Oppezzo, Chase, & Schwartz, 2010).

The SimStudent (SS) as is an artificial intelligence designed to model parts of student learning and thus stands out from the other agents in this overview as it builds a TALE around a learning model and not the opposite. The SS works by creating production rules based on examples of solved problems (see figure 5). In APLUS, the TALE part of this implementation, the student teaches simple rules of thumb for reducing algebraic equations. Both are developed by a team of researchers from Texas A & M University, Carnegie Mellon University and Oxford University (Matsuda et al. 2010).

When developing the SimStudent model of learning algebraic problem solving, Matsuda et al. (2012) found that although the same learning gains were achieved with and without the usage of SimStudent as a TA, a significantly smaller number of example problems had to be tutored.

The TALE used in this study is Guardians of History (GoH). GoH is aimed at middle school history education, especially for students in year 5, and is developed by the Educational Technology Group at Lund University.

In GoH, the student is tasked with collecting information by means of time travel and then using this information in different learning activities that the agent may observe to learn about history. The student uses a “correcting machine” to correct the task that the student has solved. In Figure 6 the activity is the timeline where the student is supposed to pair the concepts and put them on the timeline. Travels and learning activities as well as dialogue and characters are packeted in so-called missions. The missions to be performed by the student are given by the game character Professor Chronos who also provides some other guidance to the student. The missions consist of a series of activities, i.e. time travel, learning activities and verbal tests at Chronos’ office. The verbal test is simple: Chronos poses direct questions that are directly based on the learning material presented in the learning activities and the agent responds accordingly (See Figure
Figure 6: Guardians of History, timeline activity. The pieces "Émilie du Châtelet" and "built the first mirror telescope" has been put together and is to be placed on the timeline.

Figure 7: The verbal test setting. Professor Chronos asks the Time Elf: "Who or what made it possible to mass produce books at the end of the 15th century?"

Figure 8: Flow of the Guardians of History (Ternblad, 2017. Used with permission)

The flow of the play in GoH is thus quite procedural and narrative-driven (see Figure 8).

All of the beformentioned TALEs has been to variouis degrees shown to increase effort towards learning and provide learning gains. They however differ in several key aspects. One is how the learning model is implemented and another is how it is represented. BB has an explicit visual representation of Betty’s knowledge, and the TA in SF has some indication of how the TA is doing. The reciprocal LISP learning software, the learning must be inferred by the user by the gradually diminished errors the TA makes and in close resemblance to that approach is GoH where the progress is entirely driven by the narrative. SS is the only TALE with true learning algorithms, but i could be argued that both DENISE as well as BB does in fact learn as the student is providing knowledge fragments piece by piece. How the TA behaves is also a differing factor. The TA in SF ask questions and learns by example, while the TA in GoH only observes passively as the student performs learning activities. What kind of effects the behaviour of the TA and the learning model implemented has on the students’ perception, effort and performance is what is under investigation in this study.

Perception of intelligence of artificial agents

How do the students relate to the learning behaviour of TAs and how do they view the intelligence of the agents? Some research shows that 5th-graders have an understanding of the computer and the TA as not having any “real” cognitive properties, but that they engage in “play pretend” and that this behaviour is sufficient for the positive learning and metacognitive effects to take place. The students clearly held the TAs responsible for their successes and failures and even showed affection for their TAs. Schwartz et al.(2009) uses the phrase ”social metaphor”, by which”..enlist their social imagination” is meant i.e. that they are engaged in a ”play pretend”-context. Even young children (3-6 year olds) that according to other psychological measurements only have achieved limited metacognitive skills, for example theory of mind, have been shown to be able to engage in and reflect around the cognitive abilities of a TA and its learning in a TALE (Haake et al., 2015). According to Holtgraves, Ross, Weywadt, & Han (2007) humans interacting with a conversation bot clearly ascribed it human-like abilities, even if they were informed that they were interacting with a computer program.

The concept of suspension of disbelief is used in some literature (Chase, Chin, Oppezzo & Schwartz, 2009). Duffy & Zawieska (2012) discusses this concept with regards to social robotics. They argue that this is central to any kind of social setting with artificial agents, for instance robots, and that designers should embrace it. The suspension of disbelief is almost exclusively discussed in fields such as art theory and virtual reality - "immersion" - and refers to the choice of a reader, viewer etc. to voluntarily suspend judgement on the implausibility of the situation. For instance in order to disregard the limitations of the particular medium or to enjoy a magic show. A belief, unlike knowledge, does not require justification, but rather refers to the attribution of truth value of a though content. As such it can be a powerful tool to
elicit the social and cognitive effects discussed. This suspension of disbelief is thus one of the main motivating factors for learning with TAs (Schwartz et al., 2009).

But is it possible to achieve an ascription or perception of intelligence to the TA using the more fundamental human capacities of Theory of Mind? It is indeed possible to model these mind reading capabilities as a generative process that is bootstrapped by an intentional stance (Dindo, Donnarumma, Chersi & Pezzulo, 2015).

In this view, the perception of intelligence has a temporal dimension, where the initial scaffolding is tested by the perceiver over time, as a sort of ongoing experimentation of the perceiver. In that view, no suspension of disbelief regarding the learning capabilities of the TA might be required at all. At least some form of real artificial intelligence is required to be a part of the artificial agent in order to not "break the spell".

Strategies for natural language engagement such as small-talk, off-task conversation and remembering past interactions with the agents seems to be a way to increase the effects of believability and through that increasing the cognitive benefits through increased emotional connection and self-identification. The embodiment of attitudes towards the learning process and domain knowledge acquisition as well as displaying emotions can further facilitate the believability (Veletsianos & Russel (2014) who provides an excellent review of pedagogical agents in general).

In a phenomenological approach employed by Veletsianos & Miller (2008) the experiences of interactions with virtual pedagogical agents were analysed. A main theme was the oscillation in perception of virtuality and reality, never settling on one or the other. The 11 participants reported that after using a pedagogical agent for 7 weeks also used different nouns (he, she or it) to refer to the agent depending on the situation they were describing. The participants even worried about the privacy and security of the conversations even if they clearly stated that they somehow understood that the agents merely consisted of a simple AI. Their experience merged the real and the artificial without having the ability to explicitly synthesise the conflicting phenomena. Okita, Schwartz, Shibata & Tokuda (2005) supports this complex view when it comes to childrens’ attributions of intelligence to robots; the children in the experiment were less likely to view robots toy animals with larger repertoire of contingent responses and more naturalistic appearance as "pretend" and less likely to view them as "real" as opposed to robots with less naturalistic and contingent responses.

The believability of artificial agents in their role as pedagogical agents might not stem from their naturalistic embodiment. For instance, Adecock, Duggan, Nelson, & Nickel (2006) showed in a study that two learning environments, one with a embodied pedagogical agent and one with a simple helper script where perceived equally believable. Aharoni & Friðlund (2006) found, using combination of coding of facial expressions and Likert-type self-reported emotions in a mock work application interview with human and computer experimenters found that a mere label (human or computer) was enough for the participants that played the role of applicant to exhibit more interpersonal displays regardless of the actual applicant. However, some research suggests a showing of facial expressions and animations of a TA results in a larger motivation for teaching and success compared to a static, non-animated character (Bodenheimer et al., 2009).

The protégé-effect

Another theoretical concept of central importance for this study is the protégé effect i.e. that the students make greater effort towards learning in order to teach their agent than they do for themselves (Chase et al., 2009). Two studies were conducted to establish this. The first one included 62 8th-graders in two groups with a TA and "Avatar" variants of Betty’s Brain. The second study, conducted on 25 4th-graders showed that the student treated their TAs as sentient and put part of the blame of any failures on their agent.

An attempt of a breakdown of the constituting factors of this was made in the same article, where three different possible psychological phenomena where identified.

An ego-protective buffer can be generated in any peer tutoring setting. This is because the blame of failure can be put elsewhere, which shields the student in her role as a teacher from ascribing the poor performance to any internal property of her own. The success or failure of the TA can be attributed to the students role as a teacher rather than to any internal property of the student, thus protecting the ego of the student.

The adoption of an incrementalist theory about intellectual capacity in order for their efforts towards teaching their TA to be meaningful, i.e. that learning is a process of gradual increment of knowledge and skill and that neither of those is a fixed and static property.

A sense of responsibility towards their TA could be observed as the students made greater effort towards performing well in their learning tasks with TA than without even before they received any type of feedback.

These factors are further elaborated and operationalised into a questionnaire by Kirkegaard (2016). The first aspect is the incrementalist theory about knowledge i.e. the holding of the view or attitude that learning is a constructive process where knowledge is gained over time. The opposite view is to hold an ”entity theory” i.e. that knowledge is a property that is fixed or innate (Dweck, 2000). To hold the incrementalist theory about knowledge is thus an indication of the protégé-effect and the opposite holds for the entity theory. In casual, everyday language we use the word “talent” to refer to an ability that has not been trained, or has been trained to a lesser degree than normal. The opposite is to say that “effort leads to skill”. According to this, a measurement of the knowledge perspective of the protégé-effect is to ask how much effort and talent influences domain knowledge and skill.

The second aspect of the protégé-effect is the ego-protective buffer which hinges on the whether any success - or lack thereof - is ascribed to the TA. In order to shield the student from blaming herself for any failures, she must hold the TA accountable. Kirkegaard (2016) operationalises this whether a student would attribute her agent’s performance to the her own skills or teaching efforts, i.e. to measure the students attribute a successful performance to their efforts as a teacher or to attribute it to their own domain knowledge and skill.
The third aspect, the feeling of responsibility towards the TA is broken down as follows:

- The degree to which the student is learning in order to teach the TA.
- The extent to which is making an effort to teach as good as possible.
- That the student is making a greater effort when playing with the TA than without.
- How important it is for the student that the TA passes any knowledge tests.

Research Questions

If the intentional stance works as a bootstrapping device and the subsequent "experiments" by the student shows that the TA is not in fact learning would disrupt the learning process as the student no longer would perceive the TA as learning. An absence of a positive correlation between the perceived intelligence of the TA and the actual learning of the agent would support the view that "suspension of disbelief" is the driving force of the protégé-effect. However, if a positive correlation found, this will support the view that a TALE functions through an intentional stance as a bootstrapping device.

The protégé-effect is an attempt to give a theoretical outline for the beneficial learning outcomes. An aim of this study is to investigate whether a correlation between the perceived intelligence and protégé-effect is present to further develop the theoretical concept.

How much effort should be put towards the AI of the agent? This study might point towards an "AI-light" approach and a narrative-driven approach where elicitation of a suspension of disbelief is in the forefront. On the other hand, if a strong correlation between the AI, perceived intelligence and protégé-effect is found, this will entice further studies on how to elicit greater perceived intelligence through AI.

The questions under investigation in this paper are thus:

Research Question 1: Does a TA with associative learning algorithm rather than a recency learning algorithm give more perceptive effects regarding intelligence of the TA?

Research Question 2: Does a greater perceived agent intelligence correlate with a stronger protégé-effect?

Research Question 3: Is the AI of the teachable agent correlated with the learning outcome?

2 Method

Participants

For the experiment, 94 Swedish students in grades 5 and 6 in a primary school were recruited, randomly assigned one of two groups of 47 students each. No information on sex or age was available for this study, but the random assignment of half the students to one group and half to the other in every class was made in an attempt to control for any sex- or age differences.

Materials

Using the game GoH as described in the introduction with two different settings: TA_r (as in recency) and TA_a (as in associative). The recency setting makes the agent learn (and unlearn) the latest fact that it has been exposed to in the learning activities. The associative agent differs from the recency agent in three main ways: 2

1. Every time it is presented with a fact (i.e. every time the correcting machine checks a learning activity) an association is created with a certain weight. Every time this is repeated the weight increases. This is akin to a simple associative learning model.

2. At random intervals the TA uses a walkie-talkie to ask for confirmation of different facts (Figure 10).

3. In the verbal test setting, TA_r answers with the latest fact it has been exposed to while TA_a only answers if the agent has reached a threshold regarding the learned fact that is asked for. If the threshold is not met the TA asks the player for help (Figure 9). It is only in the case when the TA hasn’t been exposed to the fact at all that i doesn’t answer.

Both of the groups completed two missions. The first contained a single time travel and a concept map learning activity. The game progressed to the second mission when 80% of the learning activity was correct. The second mission consisted of 4 different time-travel locations and a time line learning activity. The mission was concluded with a verbal test for the TA at professor Chronos’ office.

Procedure

Groups of 15-20 students used GoH with two different settings (TA_r and TA_a) at two separate occasions, continuing the game in the second session. Each session lasted approx. 60min. The time between the sessions varied between one week and 2 hours.

2The TA_a was developed as a part of a project for a university course, see Bäckström, Måansson, Persson, Sakurai & Karäker Sundström (ms.)
The students were given a short introduction about the game setting and a brief introduction to the characters (Tidsalven and Professor Chronos). They were instructed to work alone but to ask for help from the teacher or from the experimenter.

A few students managed to play through both missions and get the TA through the verbal test in the first session, but most finished in the second session. Those who didn’t finish the verbal test in the second session were as mentioned above, not in the final set of participants in this study.

A questionnaire was given to the students after the completion of the two missions. The students were instructed not to assess the experience in general and to answer the questions as truthfully as the could. They were also instructed to answer the questions in the knowledge test without the aid of any notes or discussions with classmates. The questionnaire as well as the knowledge test were given in the same Google Forms form. After everyone in the class had sent their questionnaire a general group discussion of the experience was conducted. Here (among other things discussed) two questions where asked to the class: 1. Would you rather play with or without the TA and 2. Would you have played differently without the TA?

Methods of Measurement

The perceived intelligence (PI) of the agent was measured by five questions on a semantic difference-scale for measuring perceived intelligence of an robot, adapted and translated from Bartneck, Kuli & Croft (2009). (See Appendix I).

A single question about whether the student perceived, inferred and/or acknowledged any underlying learning mechanism was measured by a single question (see Appendix I). It was counted towards the PI.

Kirkengaard (2016) developed a method for measuring the protégé-effect (PE) directly through pop-up questions during gameplay. These questions are appropriate for this study and was adapted to a five question with a 5-level likert scale (See Appendix I).

The questionnaire was concluded with 10 multiple choice questions directly related to the content in Guardians of History to assess knowledge gained from the gameplay (see Appendix II).

PI, PE and the score on the knowledge test where summarised to a single score for each of the measurements.

3 Results

3 students were excluded from the dataset due to reasons of language difficulties (dyslexia or having low competence in Swedish) and another 3 because they for unknown didn’t reach the start of the second mission and were thus deemed to not be part of the experiment. An additional 3 where for unknown reasons missing from the set of questionnaire-replies, leaving N = 85 participants, with 41 in non-treatment group TA and 44 in treatment group TA (See table 1).

Two minor technical issues arose. For a handful of students the game froze when in the verbal test setting. This was solved by restarting doing another time travel and restarting the test. Another was occasional sluggish response times. None of the issues were deemed serious enough to be considered in the analysis.

All statistical analyses were performed with the statistical software R version 3.4.4 (R Core Team, 2016) at an alpha level of 0.05 if not otherwise stated. All effect sizes are interpreted against the guidelines of Cohen (1988). In the case of post hoc analyses, Bonferroni corrections were used to compensate for multiple comparisons.

The total number of participants is N = 85. A summary of the measurements are found in table 1. No significant difference between the grades where found regarding perceived intelligence (PI), protégé-effect (PE) and knowledge test performance score; they are thus treated as a single population. PI is normally distributed with non-homogenic variances, as per Shapiro-Wilk test (W = 0.97, p = .40 for TA, and W = 0.98, p = .24 for TA) and Bartlett’s test (p = .68). PE is not normally distributed, as per Shapiro-Wilk test (W = 0.95, p = .06 for TA group and W = 0.93, p < .01 for TA). The performance score of the knowledge test is not normally distributed. A summary of the distributions is seen in table 1.

Independent-samples t-test where used for normally distributed group-wise comparisons and Wilcoxon-Mann-Whitney test for non-normal distributions. A Pearson product-moment correlation coefficient where used to find correlations.

RQ1: Does the TA group perceive their TAs as more intelligent?

An independent-samples t-test was conducted to compare PI in TA and TA conditions. There were no significant difference in the PI for TA (M = 16.4, SD = 4.5) and TA (M = 17.8, SD = 4.3) conditions; t(83) = -1.32, p = .19 e.g. there was no significant difference between the TA and the TA groups with regard to the perceived intelligence as measured by the questionnaire.
RQ2: Does the degree of perceived intelligence in the TA correlate with a stronger protégé-effect?

A Pearson product-moment correlation coefficient was computed to assess the relationship between PI and PE. A strong positive correlation was found: $r = 0.64$, $p < .001$ (See figure 12). This strong correlation indicates that the students relative total score for the questionnaire for PI and PE tends to be similar.

RQ3: Does the TA lead to a better learning outcome?

A Wilcoxon-Mann-Whitney test reveals no significant difference between the groups; $W = 843$, $p = .60$. This means that there is no significant difference between the groups; neither TA$_r$- nor the TA$_a$ groups performed better in the knowledge test.

Additional findings

No significant correlation between PE and performance score in the knowledge test could be established. No additional interaction effects could be found. No other significant correlations except for the ones stated were present.

Open questions in the classroom

The two open questions in the classroom were consistent (if not unanimous). The classes all stated that 1. they’d rather play with the TA than without but that 2. they wouldn’t play differently without the TA.

4 Discussion

As the TA-groups did not differ in any significant way, a negative answer to RQ1 is indicated. The students did not perceive TA$_a$ as more intelligent. The lack of any significant difference between the groups regarding the perceived intelligence and protégé-effect is in line with my interpretation of the literature on teachable agents i.e. that the protégé effect is mainly elicited through either a narrative driven approach or by the explicitly stated role that the tutor gets as being teacher.

The strong positive correlation between perceived intelligence and protégé effect could be interpreted as either that one strongly influences the other or they strengthen each other reciprocally. Another possible explanation is that student’s general positive attitude towards the game is what is measured here. Even if they were instructed not to assess their general attitude towards the software, it is possible that the questionnaire has been interpreted that way. The data doesn’t allow for an investigation of the casual relationship between the two measurements. Clearly - even taking the questionable validity into account - the strong correlation between how the student either actively ascribe or passively perceive the TA as a thinking and learning artificial agent and the elicited protégé effect, points to the two as being worth treating in the same context and discussion. Either one has a strong influence on the other, or both are an manifestation of an underlying phenomena. Either way, it stands to reason that whether the perception of intelligence is largely a function of theory of mind, suspension of disbelief or an unfulfilled synthesis of the two (as suggested by Veletsianos & Miller (2008)), it comes along with the sought protégé-effect. Thus, it is clearly relevant and should be of great interest of any researcher and/or designer of teachable agent software to investigate the production of perceived intelligence.

A lack of correlation between the PE and performance score in the knowledge test is somewhat surprising. As the PE is constructed as a theoretical attempt to explain the positive learning outcomes from teachable agents, it might be too coarse to measured in this way. It has questionable ontological weight. That is; the result might be sensitive to false negatives. Another possible explanation is that the measurement does not reflect the PE, as noted in the previous paragraph. The PE measurement might rather reflect, or be strongly influenced, by a student’s general positive - or negative - attitude towards the software.

Questions about validity of the questionnaire instantly arise. Support of the questionnaire with the semantic difference scales as a measurement can be found in Powers & Kiesler (2006) who conducted a study on robots giving
health likeness and human likeness could be found. The same study also gives support for generalising the measurement tools to all artificial agents, as the robots in the study where virtual, animated on a computer screen. Both Powers & Kiesler (2006) as well as Bartneck, Kuli & Croft (2009) however remain humble regarding both validity and reliability. The upshot of using this kind of measurement procedure is that it limits the time the students are in a role of a participant in an experiment; they can use the majority of their time reaping the benefits from working with the TALE. Another is that since a questionnaire can easily be implemented as a part of the software, all users of the software are potential experiment participants.

There are two important assumptions made in this paper beyond what is discussed by Bartneck and colleagues (2009). One is that the translation and adaptation of the wording in the questionnaire retains the reliability and validity of the original, and that the usage of the scales, originally intended for usage with physical robots, can be generalised to be used with virtual agents.

If the perceived intelligence and protégé-effect are heavily influenced by the suspension of disbelief or the general attitude towards the setting, the introduction in the experimental protocol becomes very important. The labelling of the TA as a TA might be the single decisive factor in eliciting productive attitudes towards the agents. This is something for researchers to keep in mind.

Replies on the open questions in the classroom where consistent, the student did like the TA-character, but they also claimed that they would’ve played the same way without it. A general positive attitude towards the TALE might influence both PI and PE as measured by the questionnaire positively, i.e. that they like the setting and character and act as if they view the character as intelligent, even if they do not believe that it affects their learning behaviour.

Further research

A similar experimental protocol that is presented in this paper could be used for investigating all TALEs discussed in this paper. This would serve two purposes: to further investigate the correlations and casualities between the underlying AI, perceived intelligence, protégé-effect and learning results and to fine tune the implementation and presentation of TA. That a mere label is enough for changing the view on an artificial agent is an important first aspect to investigate.

The other TALEs presented in this paper all have similar approaches regarding how the TA is implemented: most of them have minimal actual AI. Using a similar experimental protocol as presented in this paper, different but in relevant ways similar TA software could be tested to investigate further which properties in the TALE elicit positive learning factors. One such property is the modelling of the the temporal dimension of learning. Öhlssson (2005) identifies three reasonably distinct phases in skill acquisition: initiation, mastering and optimization. In the first phase the domain is introduced and the learning is based on instructions and scaffolding. The mastering phase is largely based on getting feedback when performing satisfactory, and the optimization phase

is mostly about self-regulated learning where the importance of feedback decreases and your own memory of past achievements takes over as the main learning criterion. All of these phases comes with their own cognitive and metacognitive demands. Which means that it might require a more elaborated AI with these learning phases modelled in a longer study.

The reliability and validity of the measurements is an important step if the direction of researched proposed in this paper is to be continued. To investigate one of the measurements is to test for the general positive positive attitude towards the game. This could be done regardless of approach; both a qualitative and a quantitative setting could be utilised. In the context of this paper, a step towards validating the questionnaire would be an additional set of questions, aimed at measuring the general positive attitude towards the TALE. Likert-type questions such as "I would rather read an online article or watch instructional videos than using this software." "The information presented is more important than the story and characters," and "I would like to have the possibility to use this kind of software in all school subjects." could be used for validation when correlated with other measurements. A phenomenological approach like that of Veletsianos & Miller (2008) to verify and/or further explore the underlying mechanisms of the attitudes towards the TA and the TALE in general is another way to investigate if the questionnaire is measuring what it is supposed to measure.

A way of investigating the open questions in the classroom would be to compare the results in the experiment with a follow-up experiment where the same students play another mission in GoH without a TA and look for any difference in the behaviour in the game logs. Behavioural differences such as trial-and-error (instead of time travels), off-task activities (if non-learning games where to be included), general speed of play-through and willingness to complete the missions fully could be used to evaluate the difference in motivation and effort. This would require further development of GoH.

Concluding remarks

The main point of this theses has been twofold: to lift forth the the significance of the AI in the field of teachable virtual agents and to suggest ways to explore it. All in order to contribute to the theoretical understanding of teachable virtual agents and the development of design guidelines. Even if the underlying factors that elicit the perception of intelligence in the artificial agents remains to be disclosed, the perception as such seems to coincide with the beneficial effort inducement by teachable agents. It however seems that one only has to make an effort to behave as if an agent is learning for the the positive effects of learning by teaching is to take place.

The further development of and research on teachable virtual agents will perhaps also reveal a deeper understanding of how we humans perceive entities as intelligent. One important aspect seems to be how we choose to behave around them. What we do might be just as important as what we perceive.


Appendix I: Questionnaire

[Image of questionnaire with questions and rating scales]
Appenix II: Knowledge Test

Vem skrev Principia? *
- Newton
- Émilie du Châtelet
- Tidsalven
- Galileo Galilei

När skrevs Principia? *
- 1500-talet
- 1600-talet
- 1700-talet
- 1800-talet

Vem uppfann det första teleskopet? *
- Galileo Galilei
- Newton
- Émilie du Châtelet
- Tidsalven

När uppfanns det första teleskopet? *
- 1500-talet
- 1600-talet
- 1700-talet
- 1800-talet

Vem översatte och vidareutvecklade Principia? *
- Newton
- Émilie du Châtelet
- Tidsalven
- Galileo Galilei

Vad var det Newton upptäckte på 1600-talet? *
- Det första teleskopet
- Det första spegelteleskopet
- Ett klotexperiment
- Gravitationen

Vilken teori förknippas främst med Émilie du Châtelet? *
- Beskrivning av hur jorden rör sig kring solen
- Beskrivning av gravitationen
- Beskrivning av hur ett teleskop fungerar
- Beskrivning av hur stor rörelseenergi objekt har

Vilken bok skrev Galileo Galilei? *
- Principia
- Översättning av Principia
- De två viktigaste världsbilderna
- Att bygga ett teleskop