Challenges and approaches related to AI-driven grading in higher education: the procedural trust of students.

Author: Rozemarijn van de Leur
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT
There is a new system called 'EasyGrader' that could alter the grading process in higher education radically. The time and energy teachers spend grading open questions is increasing, and therefore the time students have to wait for their results is as well. The purpose of this study was to investigate the feasibility of an artificial intelligence-driven grading support system in higher education from the student perspective. More specifically, the goal was to look for the current level of procedural trust of students in this new system. In this context, procedural trust is defined as the degree to which a user is confident in and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid. Two major components of procedural trust have been identified: cognition-based trust and affect-based trust.

To test the hypotheses that stated that the levels of the two components of trust are sufficient, an online survey was distributed to students of higher education in Enschede. Using a Likert-scale survey, statements relating to five categories linked to the major components of trust were investigated. The results, which were analyzed using the means and modes of the answers, showed that the overall level of procedural trust in the artificial-driven support grading system is sufficient. However, some small challenges have been identified, mainly in the component of affect-based trust.

According to these findings, implementing EasyGrader in the grading process should not have major problems resulting from a lack of trust among the students.

Graduation Committee members:

Dr. Rogetzer, P.B. (First examinator)
Dr. Braun, D. (Second examinator)

Keywords
Artificial Intelligence, grading, students, support system, trust, higher education

1. INTRODUCTION

The use of Artificial intelligence (AI) has been integrated into multiple fields of work and it helps people in performing tasks associated with human intelligence (Buchanan, 2005). AI has been integrated into sectors such as health care, finance, and infrastructure (West & Allen, 2018). The use of AI in a variety of fields is being researched, including the field of education. The specific purpose of this paper is to look at an AI-driven method of supporting the grading process of open questions in the field of higher education.

The inducement for this research is the development of an AI-driven grading support system, this system is called 'EasyGrader', and the system is still in its developing phase. The system is supposed to decrease the workload for teachers by increasing the speed and decreasing the difficulty of the grading process. The system would thus advantage teachers, however, algorithms might not serve as objective and fair decision-makers, but rather reproduce biases existing within respective training data (Zoeckler, 2007).
'EasyGrader' will be a hybrid system, meaning the tool is composed of two different elements (Britannica, n.d.): AI-based grading and monitoring of the results by teachers. The result of the hybrid system should be the limitation of possible subjectiveness and biases.
This research aims to identify challenges and prerequisites related to AI-driven grading support in higher education. Specifically, this research investigates the prerequisite: procedural trust of students in an AI-driven grading system in higher education. The goal is to find out if students are ready for such a change. If the study concludes that there is a lack of trust in this system, because of for example an increasing chance of biases, the developer could experience resistance, which would challenge the implementation of a change (Kotter & Schlesinger, 1989). If the study concludes that the students do trust such a system, EasyGrader can use this as an argument for collaboration with educational institutions. The grading process could be changed radically.

The University of Twente (UT) is looking into the possibility to implement the support system. This research helps to assess the feasibility for the UT for the appliance of such a system, from a student perspective. The tool is meant for higher education in general; higher education in the Netherlands comprises higher professional education (HBO) and university education (WO) (Ministerie van Onderwijs, Cultuur en Wetenschap, 2020). Due to this classification, this research uses the perspective of students of higher education in Enschede:

- Students from the UT (the only WO institution in Enschede)
- Students from Saxion (one of the two HBO institutions in Enschede)
- Students from Artez (one of the two HBO institutions in Enschede, focussing specifically on education in the field of art)

1.1 Research question

As mentioned before the central aim of this paper is to research the level of procedural trust of students in the AI-driven support grading system.

The problem of a low level of procedural trust is the possible resistance that could result from this when such a system would be implemented without the consideration of this challenge.

This goal and problem statement have led to the following research question:

What is the current level of procedural trust of higher education students in the AI-driven grading support system 'EasyGrader' in Enschede?

In the following, the topics of AI, procedural trust, and the trust in AI systems will be introduced and explained in order to provide the necessary information to fully understand this research and to clarify to which definitions this research is restricted. Furthermore, the model of human-computer trust components created by Madsen and Gregor (2000) is introduced. This framework was used as the base of this research. Hypotheses are formulated with the use of this framework. To test these hypotheses, statements were formulated and distributed with the use of an online survey. In Section 3, this method of research is displayed. Following the methodology section, the results of the survey will be discussed in Section 4. Based on these results a clear conclusion on the current state of the procedural trust of the students will be defined. Lastly, some limitations of this study will be explained.

1.2 Research contribution

There has already been a lot of research on (procedural) trust and AI systems, e.g. (Banavar, 2016) and (Morse et al. 2021). Many funds have been received for AI research (Anjila, 2021). Furthermore, the implementation of AI in the specific sector of education has been researched (Goksel & Bozkurt, 2019), (Chen et al., 2020). An increase in the research on AI in education has been identified, which indicates an increase in the importance and demand for the implementation of AI technologies in education (Chen et al., 2020). However, research on AI in education has been mainly focused on the learning processes and not on the grading processes. The implementation of AI in the grading process is therefore in need of more research. This research contributes to narrowing this research gap.

Before a higher education institution will implement the system, it is important to know the possible consequences of the system to make a fitting strategy. This strategy is needed to tackle any challenges. One possible challenge of change could be resistance, of which lack of trust is a possible cause (Kotter & Schlesinger, 1989). As it is in the best interest of educational institutions to avoid student resistance, this research is relevant for the implementation of AI-driven grading support systems in practice as well. Resistance of students could consequence in a declining number of students, which in turn decreases the monetary income of the educational institution.

From a student's perspective, trust is extremely important for student performance (Cheng et al., 2017), which is why it is important to consider the trust of the students for their benefit.

2. THEORETICAL FRAMEWORK

As there is no general consensus on both the definition of artificial intelligence and trust the theoretical framework of this report aims to reach a shared perspective with the reader on the definitions of these subjects. Furthermore, hypotheses were formulated based on the Model of human-computer trust components which will be explained in Section 2.3. Lastly, a literature review has been displayed in Section 2.5.

2.1 Artificial intelligence (AI)

Artificial intelligence (AI) is a term on which multiple definitions have been formulated. Some of these definitions are:

- "The ability of a digital computer or a computer-controlled robot to perform tasks commonly associated with intelligent beings" (Copeland, 2021).
- "The simulation of human intelligence in machines that are programmed to link human beings and mimic their actions" (Frankenfield, 2021).
- "The simulation of human intelligence processes by machines, especially computer systems" (Burns et al., 2022).

For this research, the definition of AI is restricted to the definition of AI systems, for the reason that this research is about a specific AI system (the grading support system). There are multiple definitions for the systems also, and most of them can be categorized into four categories: systems that think like humans, systems that act like humans, systems that think rationally, and systems that act rationally (Kok et al., 2002). The AI-driven grading support system should think and act like humans in grading the open question, which is along the lines of the definition in two of the four categories.

2.2 (Procedural) trust

Trust in general is a highly researched topic without one cohesive definition which is adopted by all. Multiple studies have given multiple classifications and forms of 'trust', e.g. (Mcknight & Chervany, 1996), (Höhmann et al., 2005), (Pytlikzillig & Kimbrough, 2015). Despite the divergence in conceptualizations of trust, a majority of authors agree that trust is a psychological state (Li & Betts, 2003). Furthermore, trust is often seen as a choice; when we choose to trust or not to trust we are making a decision. This research restricts itself to the definition: trust is a choice to place one's confidence in others (Li & Betts, 2003).

Procedural trust is the trust in procedures or other systems that decrease the vulnerability of the potential trustor, enabling action in the absence of other forms of trust (Stern et al., 2014). This type of trust is the best type to do research upon for this thesis as the preferred outcome is to get insight into the trust student have in the grading process when AI-driven grading is implemented.

Specifically, this research is aimed to study the procedural trust in an AI system. The procedural trust in an AI decision aid is also called 'human-computer trust', human-computer trust is "the extent to which a user is confident in and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid." (Madsen & Gregor, 2000). For this form of trust, an individual would thus choose to place confidence in an artificially intelligent decision aid.

2.3 Framework of human-computer trust

There are multiple theories on how to examine trust and how to examine human-computer trust. A widely used theory is the theory of Jian et al. (2000), who identified that the five words most related to trust between human and automated systems were trustworthy, loyalty, reliability, honor, and familiarity. Another theory developed to measure human-computer trust is that of Madsen and Gregor (2000). This study deals with intelligent systems which are designed to aid decision-making. As this research is aimed to analyze the trust in an intelligent system aimed to aid decision-making, this is the most fitting theory to use. This study identifies the relationship between perceived understandability, perceived technical competence, perceived reliability, personal attachment, and faith in cognition-based trust and affect-based trust. The last two aspects are the two main components of human-computer trust. These two components of trust have been identified before (McAllister, 1995) in which cognition-based trust was defined as the rational evaluation of an individual and affect-based trust was defined as the emotional attachment. Madsen & Gregor (2000) expanded this work with factors identified to affect the trust in AI systems. A conceptual framework illustrating the relationships between the different components can be seen in Figure 1.

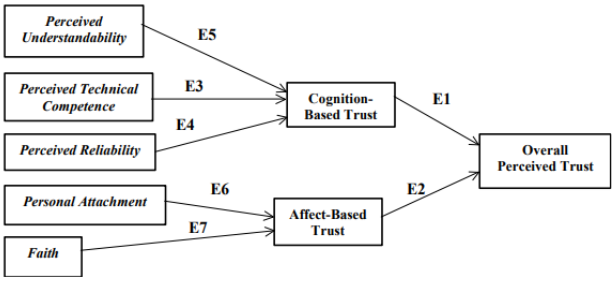

Figure 1 Model of human-computer trust components (Source: Madsen & Gregor 2000)

2.4 Hypotheses

To research the perceived trustworthiness of AI-driven grading support systems two main hypotheses were set up on the components of the human-computer trust model. These main hypotheses have five sub-hypotheses in total.

H1: The cognition-based trust in an AI-driven grading support system is sufficient. This trust is sufficient when hypotheses H1a, H1b, and H1c are accepted.

- H1a: The AI-driven grading support system is perceived as reliable. Reliability means consistency when the use of something is repeated (Ribana Hategan, 2020).
- H1b: The AI-driven grading support system is perceived as technically competent. Technical competence means that the system performs the tasks accurately and correctly based on the information that is input (Madsen & Gregor, 2000).
- H1c: The AI-driven grading support system is perceived as understandable. Understandability means that the human supervisor or observer can form a mental model and predict future system behavior (Madsen & Gregor, 2000).

H2: The affect-based trust of an AI-driven grading support system is sufficient. This trust is sufficient when hypotheses H2a and H2b are accepted.

- H2a: The students have faith in an AI-driven grading support system. Faith means that the user has faith in the future ability of the system to perform, even in

situations in which it is untried (Madsen & Gregor, 2000).

- H2b: There is a personal attachment to the AI-driven grading support system. Personal attachment means that the user finds using the system agreeable and it suits their taste and love, meaning that the user has a strong preference for the system, is partial to using it, and has an attachment to it (Madsen and Gregor, 2000).

2.5 Literature review

Change can be frightening, especially when a long-standing way of doing things is changed (Weiner & Bornstein, 2009). Even though there are multiple types of grading practices (McMillan et al., 2010), the grading process in higher education has mainly been manual, as the technology has only recently been developed. On top of this, robots tend to evoke the emotion of fear as well (Glikson & Woolley, 2020). Robots use AI and are often related to AI (Berezina et al., 2019). Even though the basis for personal relationships is often faith, the human-computer trust often works in the reversed way. Faith is an aspect that could benefit trust but often develops later with human-computer trust (Hoff & Bashir, 2014).

A comparable system to the system researched in this report has been used in a study in Florida. This system was used to grade the essays of students. The only weakness reported was related to technical issues (Burstein et al., 2021). Technical issues may also be a challenge for the system subject to this research, and the perception of students on the technical aspect. This feedback in Burnstein et al. (2021), was however reported by teachers. The opinion of students was not taken into account in his research. An advantage for students which was identified is the immediate feedback, is the fast grading process This benefit is also there for students for the system central in this report.

Errors of AI systems that are visible to the user(s) can affect trust in a way that is difficult to repair (Glikson & Woolley, 2020). An example of a visible error of the grading support system which might occur is a falsely graded outcome. This again shows the importance of perceived technical competence on trust. It also shows the possibility for trust to decrease after the implementation of an AI system because of malfunctions (Glikson & Woolley, 2020).

Earlier research has also shown that people tend to trust human decision-making more than algorithm decision-making in tasks that involve human skills, such as work evaluations (Glikson & Woolley, 2020). This is again related to the technical competence of the system. For the grading support system, this would mean that participants would trust teachers more to evaluate their answers than in the AI-support system.

In Spain, research was conducted in which attitudes towards AI were analyzed. It was discovered that people were overall positive towards new AI technologies being developed, however, with the requirement that this would benefit society. Negative attitudes came from a fear that the innovations would harm society (e.g. decrease the number of jobs) (Albarrán Lozano et al., 2021). The research of Albarrán Lozano et al. (2021) is relevant as it gives evidence that the attitudes of people towards new AI developments can be more positive when positive effects are seen. The innovations can also give a feeling of fear or threat once the belief is there that innovation may harm society/people (Albarrán Lozano et al., 2021). Personal attachment and perceived reliability could either make or break the implementation of an innovation.

3. METHODOLOGY

This section describes the method which was used to test the hypotheses and argues why this method was used and how it was used. The second paragraph displays details of the research sample e.g. the average age and the distribution of their gender.

3.1 Research method

The research method which was used is the conduction of an online survey. This online survey could be filled in by participants either through a mobile device, computer, or tablet. The survey was active for ten days from May 12th, till May 22th, 2022.

The framework introduced in Section 2.3 is intended to be used for questionnaires, which is why a survey is the most fitting method for this research. With the components identified to influence cognition-based trust and affection-based trust, 25 corresponding items were constructed by Madsen & Gregor (2000). These items are statements. Each component (perceived understandability, perceived technical competence, perceived reliability, personal attachment, and faith) has an item battery consisting of either four or five items/ statements which test each component. The statements are adjusted to fit this research and measure the specific grading-support system. A full overview of the statements used in the online survey can be found in Appendix 3. The survey was designed using the platform 'Qualtrics XM' which was presented by the BMS faculty as the best choice because of its functionalities, security, and privacy measures (BMS Lab University of Twente, 2016).

The statements are Likert scale statements, with the options 'strongly disagree', 'disagree', 'neither disagree nor agree', 'agree', and 'strongly agree' (an uneven number for students to be able to choose a neutral option). The scales are 1= strongly disagree, 2= disagree, 3= neutral, 4=agree & 5= strongly agree.

Lastly, demographic questions are asked to get insight into the participants': -age, -gender, and -perceived technical knowledge. With these questions, the sample population of the survey was able to be analyzed.

It is important to mention that the participants were asked to answer based on their opinions and attitudes towards the statements. They did not get additional information other than what is shown in Appendix 1 . An example of what is meant by this: when assessing if they could rely on the system to function properly, they were not given information about possible errors the system already has or could produce.

3.2 Research sample

The participants were acquired by: the distribution of the link to the survey on social media (LinkedIn, Instagram, and WhatsApp Groups), and through distributing flyers at Saxion, the UT, and Artez which contained a QR code leading to the survey.

In order to fully participate in the survey, the participants needed to meet some requirements. First of all, they had to consent with participating in the research after reading the opening statement. Secondly, the students have to study at either the UT, Saxion Enschede or Artez Enschede, as this is the population this research is aimed at.

To assess if students met the requirements two questions were placed at the beginning of the survey but after the opening statement. 'Do you consent to participate in this study?' and 'Do you study at one of these educational institutions: the UT, Saxion Enschede, or Artez Enschede?'. If a participant answered 'no' to one of these questions, he/she/them could not continue filling in the survey.

A total of 77 participants who participated met the requirements. The average age of the participants was 21.88 years, no participants were older than 26. 44.9% of the participants identified themselves as male, and 55.1% of the participants identified themselves as female. Furthermore, participants had to rank their technical knowledge on a rank from 1 to 5 with 1 meaning no technical knowledge and 5 meaning the highest level of technical knowledge, the average level was 3.15.

In Appendix 2, a detailed overview of the research sample can be found.

4. RESULTS

In order to analyze the results properly, the program SPSS was used. The reasons for the selection of this program were: (1) the author of this report is specialized in using this program, (2) it provides many ways to examine data (College & Flynn, 2003) (3) 'Qualtrics', the program used to distribute the surveys of this research, can be linked to SPSS, making it possible to directly and precisely transfer the data from one program to the other.

To analyze the results the Likert scale was given 'scores'. These scores are: 'strongly disagree' = 1, 'disagree'= 2, 'neither disagree nor agree' = 3, 'agree' = 4, and 'strongly agree' = 5.

All of the questions referred to in the results can be found in Appendix 3.

4.1 Analysis preparation

The mean is said not to be meaningful when a Likert scale is used e.g. what is the average between agreeing and strongly agreeing? (University of St Andrews, 2022). Further research has proven that means can be used with Likert-scale questions, and have even recommended using this. (Sullivan & Artino, 2013). The analysis, therefore, did analyze means as the non-numerical scales were transformed to numerical scores, which is shown in Section 4. Modes were also included to show which score was given for each question by the majority of the participants. A full overview of the results is shown in Figure 2.

4.2 Cognition-based trust

4.2.1 Perceived reliability

For the first aspect, perceived reliability (the variable is abbreviated to 'R' in SPSS), it was measured if the participants perceived the system to be reliable. Statements R1 to R5 relate to perceived reliability. The sub-hypothesis related to perceived reliability is hypothesis H1c i.e. "The AI-driven grading support system is perceived as reliable.". This hypothesis will be rejected when the mean of the perceived reliability ≤ 3 and/or when the mode is <3. This limit was chosen as all statements were formulated in a positive relation to the variable, e.g. a high score on statements R1 to R5 indicates good perceived reliability. Three is the exact boundary, meaning neither disagree nor agree. This being said, a score lower than three indicates a 'negative' score in relation to the variable. As all hypotheses were formulated 'positively' in relation to the variable, the negative score would mean rejection of the hypothesis.
The hypothesis will be accepted when the mean of the perceived reliability is > 3 and/or when the mode is ≥ 3.

In the results, it can be seen that the mode for all questions is 4/ 'agree'. For some questions, this is more convincing than for others, e.g. the questions about consistency (R3 & R5), have a higher mean than the question on the quality of the advice the system produces (R1). However, as both the mode and the mean of perceived reliability are above three, the perceived reliability of the system is good, and hypothesis H1c is therefore accepted.

4.2.2 Perceived technical competence

Secondly, the perceived technical competence (the variable is abbreviated to 'T' in SPSS) of the participants was analyzed. The hypothesis related to perceived technical competence is hypothesis H1b i.e. "The AI-driven grading support system is perceived as technically competent.". This hypothesis will be rejected when the mean of technical competence is ≤ 3 and/or when the mode is <3 and accepted when the mean of technical competence is > 3 and/or when the mode is ≥ 3. for the same reason as was formulated in Section 4.2.1. Statements T1 to T4 relate to technical competence.

Statement T3, "The system correctly reviews the answers to open questions I enter..", has a relatively low mean of the component T. The mode for this question was the lowest mode of the component, 3, indicating that a majority of participants neither disagree nor agree with the statement. This is concerning as it shows a lack of trust in the technical competence of the grading process.

Statement T2 is also notable, as this statement scored the lowest mean (2,88) of all statements related to technical competence. The statement was "The advice the system produces will be as good as that which a highly competent person (a professor for example) could produce.". This is concerning as it shows a lack of trust in the technical competence of the grading process of the system pertaining to the competence of a human being in the grading process.

Nevertheless, with a mean of 3.3 and a mode of 4, participants perceived the technical competence of the system as 'good'. Hypothesis H1b has been accepted.

4.2.3 Perceived understandability

The last component of cognition-based trust is perceived understandability (the variable is abbreviated to 'U' in SPSS). With this variable, it was analyzed if participants found the system understandable. Regarding perceived understandability, hypothesis H1c was formulated, i.e. "The AI-driven grading support system is perceived as understandable.". In consistency with the variables analyzed before, this hypothesis will be rejected when the mean of the perceived understandability ≤ 3 and/or when the mode is <3 and accepted when the mean of the perceived understandability > 3 and/or when the mode is ≥ 3. Questions U1 to U4 relate to the perceived understandability.

In the results, it can be seen that for all questions related to perceived understandability a majority of the participants agreed with each individual statement. The mode, therefore, is 'agree' for questions U1 to U4 and the perceived understandability in general. The mean also shows more people agree than disagree. There were no remarkable differences in the results for the different statements. The means and modes all exceeded three, meaning hypothesis H1c was accepted.

4.2.4 Hypothesis 1

The first three variables discussed in the results all come together in the first main hypothesis, H1 i.e. "The cognition-based trust in an AI-driven grading support system is sufficient." This trust is sufficient when hypotheses H1a, H1b, and H1c are accepted.

As a majority of the participants agreed on statements R1 to R5, the system can be said to be perceived as reliable by the students from Enschede. The hypothesis of this component (H1a) was accepted.

Furthermore, technical competence was analyzed. This aspect has proven to be important as once there is a visible error in the system, trust is hard to repair (Glikson & Woolley, 2020). Challenges have been identified in this section. Participants showed concern about the quality of advice the system produces relative to what a competent person could produce. Even though concerns were raised, hypothesis H1b was accepted based on the mean and mode of technical competence.

Lastly, the perceived understandability of the system was analyzed. As the understanding of the tasks of an AI system has a positive effect on procedural trust (Hoff & Bashir, 2014), this aspect is again considered to be important. The mode for all related statements was 'agree', which indicates that there are currently no problems with the understandability of the AI-driven grading support system. Hypothesis H1c has been accepted.

As hypotheses H1a, H1b, and H1c have been accepted, hypothesis H1 is accepted. The cognition-based trust in the AI-driven grading support system 'EasyGrader', therefore, is sufficient.

4.3 Affect-based trust

4.3.1 Faith

The first component of 'affect-based trust' is faith (the variable is abbreviated to 'F' in SPSS). The hypothesis related to perceived technical competence is H2a i.e. "The students have faith in AI-driven grading support system.". This hypothesis will be rejected when the mean of faith is ≤ 3 and/or when the mode is <3 and accepted when the mean of faith > 3 and/or when the mode is ≥ 3. for the same reason as was formulated in Section 4.2.1. Questions F1 to F4 were related to faith.

In component 'faith', statement F3 is immediately notable. This is the statement with the lowest mean and mode of all statements (2.81 & 2). The statement is "When the system gives unusual advice I am confident that the advice is correct".

In contrast to statement F3, the other statements indicate that there is faith in the AI-driven support grading system, as they all have a mode of four and a mean above three. It might be that the participants indicated that there is faith, but not as good to trust unusual advice above their own reasoning.

As the overall mean of faith is above three (3.06), which also counts for the mode (4), hypothesis H2a is accepted.

4.3.2 Personal attachment

Lastly, the results of personal attachment (the variable is abbreviated to 'P' in SPSS) were analyzed. The hypothesis related to perceived technical competence is H2b i.e. "There is a personal attachment to the AI-driven grading support system.". This hypothesis will be rejected when the mean of personal attachment is ≤ 3 and/or when the mode is <3 and accepted when the mean of personal attachment is > 3 and/or when the mode is ≥ 3. The questions related to personal attachment were P1 to P5.

The results have shown the mode to be 'agree' for all statements but one. For statement P2 'I feel a sense of attachment to the system' the median was shown to be 'neither disagree not agree'. The mean for this statement is 2.82, which is relatively low. This indicates that the participants did not feel a sense of attachment to the system.

Nevertheless, the other statements did indicate a slight sense of personal attachment to the system. As the mean was >3 and the mode was also >3, hypothesis H2b has been accepted.

4.3.3 Hypothesis 2

"The affect-based trust of an AI-driven grading support system is sufficient.", is main hypothesis 2. This trust is sufficient when hypotheses H2a and H2b are accepted.

As we saw in Section 4.3.2 there was a concern surrounding the faith in the system when it would produce unusual advice. Besides this, there was a high enough level of faith to accept the related hypothesis H2a.

In the next section on personal attachment, it became clear that the students did not feel a sense of attachment to the system. The other statements, however, scored above the mean of three and above the mode of three. This made P overall have a 'positive' score. The hypothesis related to this variable, H2b, was therefore accepted.

As both H2a & H2b were accepted, H2 has been accepted and it can be said that the affect-based trust of the AI-driven grading support system is sufficient.

It is noteworthy to mention that the components related to hypothesis two both had lower mean scores than the components related to hypothesis one. This indicates more cognition-based trust than affect-based trust towards the AI-driven support grading system.

	Mean	Standard deviation	Mode
R1	3.12	0.822	4
R2	3.45	0.784	4
R3	3.73	0.767	4
R4	3.36	0.805	4
R5	3.67	0.75	4
R	3.46	0.62	4
T1	3.45	0.8	4
T2	2.88	1.032	4
T3	3.32	0.781	3
T4	3.6	0.744	4
T	3.3	0.633	4
U1	3.4	0.917	4
U2	3.51	0.849	4
U3	3.55	0.832	4
U4	3.49	0.818	4
U	3.49	0.72	4
F1	3.03	0.882	4
F2	3.28	0.864	4
F3	2.81	0.968	2
F4	3.22	0.907	4
F	3.06	0.788	4
P1	2.92	1.016	4
P2	2.82	1.003	3
P3	3.22	0.921	4
P4	3.22	0.962	4
P5	3.13	1.011	4
P	3.06	0.828	4

Figure 2 The survey results of all individual statements

	Mean	Standard deviation	Mode
R	3.46	0.62	4
T	3.3	0.633	4
U	3.49	0.72	4
F	3.06	0.788	4
P	3.06	0.828	4

Figure 3 The survey results of the trust components

	Mean	Standard deviation	Mode
Cognition-based trust	3.42	0.563	4
Affection-based trust	3.06	0.776	4

Figure 4 The survey results of the trust categories

5. DISCUSSION

The acceptance of the hypotheses shows that the level of both cognition-based trust as well as affect-based trust is already sufficient. This indicates a possibility to implement the system without trouble regarding trust when students are given the same amount of information to as this study did.

Some challenges have been identified in individual statements, e.g. T2, F3, P1, and P2. Even though these concerns did not have a high level of effect on the end conclusions as the hypotheses were still accepted, they can be taken into account during the implementation of EasyGrader.

For cognition-based trust, the largest challenge is the trust in technical competence. Students did not agree on this AI-driven system to be just as competent as a competent person would be in grading their open questions. This is in line with earlier research which has shown proof that people tend to trust human decision-making more than algorithm decision-making in tasks that involve human skills (Glikson & Woolley, 2020), such as grading open exam questions.

A way to increase the trust related to technical competence is to avoid errors. As identified before, once an error has been visible the trust is hard to repair (Glikson & Woolley, 2020). As the trust related to the technical competence of the system is already relatively low, it should be avoided to be damaged even more. A way to steer clear of errors is to implement bias detection and mitigation capabilities on the AI-driven grading support system (Rossi, 2018).

For the affect-based trust, the largest challenge would be to increase the faith students have in the system. A possible explanation for the lack of faith is that human-computer trust often works in the opposite way personal relationships work (Hoff & Bashir, 2014). The faith might develop later provided that there is satisfaction with the system. The sense of personal attachment could also be improved as a consequence.

However, some research has shown that low trust can harm innovation, but very high trust does the same (Bidault & Castello, 2010). The combination of a sufficient level of trust with some concerns might therefore be positive.

As the overall level of human-computer trust has shown to be sufficient, this suggests that there will be no major issues during and after the implementation of the system of EasyGrader which can be linked to a lack of trust. This means that there should be no resistance, no declining number of students, and no decreasing student performance because of a lack of trust in the grading process.

EasyGrader could use the sufficient level of trust in the system as an argument for collaboration with educational institutions.

6. LIMITATIONS

This study has some limitations. First of all, due to a lack of research experience of the author, some biases may have been raised with the data implementation.

Furthermore, an impact limitation worthy to mention is the strong regional focus of this research. The study was focused on higher education in Enschede. Because of this, the results may not be applicable for other regions in the Netherlands, as well as to the rest of the continent or world.

Lastly, students who filled in the survey did not work with/ see the current version of the system. The survey was filled in based on a description of the AI grading support system. Perceptions of how the system works might therefore differ from the actual working of the system.

7. CONCLUSION

Throughout this work, it has been argued that human-computer trust is an important factor to research before the implementation of a new AI-grading support system will be set in motion. The reasons for this are the (1) possibility of resistance and (2) the effect of trust on student performance.

The aim of the research was to look at the current level of human-computer trust in an AI-driven support grading system for higher education. Throughout the research, the goal was to identify possible challenges, which could occur when the technology would be implemented right away with a brief explanation, e.g. the explanation used in the survey of this research. The development of a fitting strategy will be easier with knowledge of possible challenges.

In the literature review, some suggestions were recognized that the current state of trust might not be optimal. This mainly had to do with how new the system was.

A framework by Madsen and Gregor (2000) identified two categories of human-computer trust. These aspects in their turn had a total of five components. These levels of these components were tested using an online survey in which Likert-scale questions were displayed.

The outcomes of this work are relevant as the effects of no trust in the system might now be avoided. During the implementation of the system, educational institutions could focus on the areas in which students have shown the most concerns. For students, the research has been highly relevant as they have been given the opportunity to express their concerns. People react more strongly when the procedure does not give them a voice, even though it might affect them (van den Bos & van Prooijen, 2001). Apart from the emotion of the students who have been given a voice, educational institutions might act upon their voice.

A sufficient level of trust in the AI-driven support grading system (EasyGrader) was identified, which indicates a low chance of resistance and a low chance of a negative effect on student performance. Some areas that should be handled with care, have also been recognized. These ‘areas’ are components that scored relatively low in trust in the survey. These components were mostly linked to the category of human-computer trust: affect-based trust.

8. REFERENCES

Albarrán Lozano, I., Molina, J. M., & Gijón, C. (2021). Perception of Artificial Intelligence in Spain. *Telematics and Informatics*, *63*(101672), 101672. https://doi.org/10.1016/j.tele.2021.101672

Banavar, G. (2016, November 29). *What It Will Take for Us to Trust AI*. Harvard Business Review. https://hbr.org/2016/11/what-it-will-take-for-us-to-trust-ai . Retrieved on June 6th 2022.

Berezina, K., Ciftci, O., & Cobanoglu, C. (2019). *Robots, Artificial Intelligence And Service Automation In Travel, Tourism And Hospitality*. Emerald Group Publ. https://www.emerald.com/insight/content/doi/10.1108/978-1-78756-687-320191010/full/html

Bidault, F., & Castello, A. (2010, May 10). *Why Too Much Trust Is Death to Innovation*. MIT Sloan Management Review. https://sloanreview.mit.edu/article/why-too-much-trust-is-death-to innovation/#:~:text=While%20personality%20conflicts%20hinder%20innovation,to%20implement%20jointly%20developed%20ideas. Retrieved on May 25th 2022.

BMS Lab University of Twente. (2016). *Survey Software | BMS - BMS Lab*. Universiteit Twente; BMS Lab. https://www.utwente.nl/en/bms/datalab/datacollection/surveysoftware/ . Retrieved on May 17th 2022.

Buchanan, B. G. (2005). A (Very) Brief History of Artificial Intelligence. *AI Magazine*, *26*(4), 53–53. https://doi.org/10.1609/aimag.v26i4.1848

Burns, E., Laskowski, N., & Tucci, L. (2022, February). *What is Artificial Intelligence (AI)? - AI Definition and How it Works*. SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence . Retrieved on April 6th 2022.

Burstein, J., Chodorow, M., & Leacock, C. (2021). Automated Essay Evaluation: The Criterion Online Writing Service. *AI Magazine*, *25*(3), 27–27. https://doi.org/10.1609/aimag.v25i3.1774

Cameron, E., & Green, M. (2015). *Making Sense of Change Management*. Kogan Page Publishers.

Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, *1*, 100002. https://doi.org/10.1016/j.caeai.2020.100002

Cheng, X., Fu, S., Han, Y., & Zarifis, A. (2017). Investigating the individual trust and school performance in semi-virtual collaboration groups. *Information Technology & People*, *30*(3), 691–707. https://doi.org/10.1108/itp-01-2016-0024

College, B., & Flynn, D. (2003). *Student Guide to SPSS*. https://faculty.ksu.edu.sa/sites/default/files/student_user_guide_for_spss.pdf . Retrieved on June 5th 2022.

Copeland, B. J. (2021). artificial intelligence | Definition, Examples, and Applications. In *Encyclopædia Britannica*. https://www.britannica.com/technology/artificial-intelligence . Retrieved on April 4th 2022.

Frankenfield, J. (2021, March 8). *How Artificial Intelligence Works*. Investopedia. https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp#:%7E:text=Artificial%20intelligence%20(AI)%20refers%20to . Retrieved on April 6th 2022.

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2). https://doi.org/10.5465/annals.2018.0057

Goksel, N., & Bozkurt, A. (2019). Artificial Intelligence in Education. *Handbook of Research on Learning in the Age of Transhumanism*, 224–236. https://doi.org/10.4018/978-1-5225-8431-5.ch014

Hoff, K. A., & Bashir, M. (2014). Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Höhmann, H.-H., & Welter, F. (2005). *Trust and entrepreneurship: A west-east perspective* (pp. 7–16). Edward Elgar Publishing.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/s15327566ijce0401_04

Kok, J., Boers, E., Kosters, W., Van Der Putten, P., & Poel, M. (2002). *ARTIFICIAL INTELLIGENCE -Artificial Intelligence: Definition, Trends, Techniques and Cases - ARTIFICIAL INTELLIGENCE: DEFINITION, TRENDS, TECHNIQUES, AND CASES*.

Kotter, J., & Schlesinger, L. (2008, August). *Choosing strategies for change*. Harvard Business Review. https://hbr.org/2008/07/choosing-strategies-for-change . Retrieved on May 10th 2022.

Li, F., & Betts, S. (2003). Trust: What It Is And What It Is Not. *International Business & Economics Research Journal*, *2*(7), 67–75. https://doi.org/10.19030/iber.v2i7.3825

Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. *Proceedings of the 11 Th Australasian Conference on Information Systems*, 6–8.

McAllister, D. J. (1995). Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal*, *38*(1), 24–59. https://doi.org/10.5465/256727

Mcknight, D., & Chervany, N. (1996). *THE MEANINGS OF TRUST*.

McMillan, James. H., Myran, S., & Workman, D. (2010). Elementary Teachers' Classroom Assessment and Grading Practices. *The Journal of Educational Research*, *95*(4), 203–213. https://doi.org/https://doi.org/10.1080/00220670209596593

Morse, L., Teodorescu, M. H. M., Awwad, Y., & Kane, G. C. (2021). Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-021-04939-5

Pytlikzillig, L., & Kimbrough, C. (2015). *Consensus on Conceptualizations and Definitions of Trust: Are We There Yet? A longitudinal and experimental study of the impact of knowledge on the bases of institutional trust View project Central Great Plains Climate Education Partnership View project*. https://doi.org/10.13140/RG.2.1.1365.5205

Ribana Hategan. (2020, December 7). *Vibration testing equipment*. Vibration Testing Equipment. https://www.etssolution-asia.com/blog/definition-of-reliability-and-validity . Retrieved on May 9th 2022.

Rossi, F. (2018). BUILDING TRUST IN ARTIFICIAL INTELLIGENCE. *Journal of International Affairs*, *72*(1), 127–134. https://www.jstor.org/stable/26588348

Stern, M. J., & Coleman, K. J. (2014). The Multidimensionality of Trust: Applications in Collaborative Natural Resource Management. *Society & Natural Resources*, *28*(2), 117–132. https://doi.org/10.1080/08941920.2014.945062

Sullivan, G. M., & Artino, A. R. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/jgme-5-4-18

University of St Andrews. (2022). St-Andrews.ac.uk. https://www.st-andrews.ac.uk/ . Retrieved on May 24th 2022.

van den Bos, K., & van Prooijen, J.-W. (2001). Referent Cognitions Theory: The role of closeness of reference points in the psychology of voice. *Journal of Personality and Social Psychology*, *81*(4), 616–626. https://doi.org/10.1037/0022-3514.81.4.616

Weiner, I. B., & Bornstein, R. F. (2009). Principles of Psychotherapy. (p. 172). John Wiley & Sons inc.

West, D. M., & Allen, J. R. (2018, April 24). *How artificial intelligence is transforming the world*. Brookings; Brookings. https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/#_edn2 . June 13th 2022.

Zoeckler, L. (2007). Moral Aspects of Grading: A Study of High School English Teachers' Perceptions. *Undefined*. https://www.semanticscholar.org/paper/Moral-Aspects-of-Grading%3A-A-Study-of-High-School-Zoeckler/0e390c1957cedb4978b71da95d8f7037634ce801

9. APPENDIX

Appendix 1: Information gives beforehand

'*EasyGrader*' is a **new artificial intelligence system**, which is meant to **grade open questions** in higher education. It will **advice** a grade teachers on the grade, they should check the outcomes.
The system will use **training data**.

Please response to the statements with your **assumptions**. There is **no 'right or wrong'**

Appendix 2: Research sample

sn

➔ **Descriptives**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
With 1 being the lowest level and 5 the highest, how would you rate your current technological knowledge?	78	1	5	3.15	.774
What gender do you identify as?	78	1	2	1.55	.501
What is your age?	78	1	9	4.88	1.299
Valid N (listwise)	78				

With 1 being the lowest level and 5 the highest, how would you rate your current technological knowledge?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	1.3	1.3	1.3
	2	12	15.4	15.4	16.7
	3	42	53.8	53.8	70.5
	4	20	25.6	25.6	96.2
	5	3	3.8	3.8	100.0
	Total	78	100.0	100.0	

Frequency Table

What is your age?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18 years or younger	1	1.3	1.3	1.3
	19	1	1.3	1.3	2.6
	20	8	10.3	10.3	12.8
	21	17	21.8	21.8	34.6
	22	29	37.2	37.2	71.8
	23	16	20.5	20.5	92.3
	24	4	5.1	5.1	97.4
	25	1	1.3	1.3	98.7
	26	1	1.3	1.3	100.0
	Total	78	100.0	100.0	

What gender do you identify as?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	35	44.9	44.9	44.9
	Female	43	55.1	55.1	100.0
	Total	78	100.0	100.0	

Appendix 3: Survey statements

1. Perceived Reliability

R1 - The system will always provide the advice teachers require to make their decisions. R2 - The system will perform reliably. R3 - The system responds the same way under the same conditions at different times. R4 - I can rely on the system to function properly. R5 - The system analyzes problems consistently.

2. Perceived Technical Competence

T1 - The system uses appropriate methods to reach decisions. T2 - The advice the system produces will be as good as that which a highly competent person (a professor for example) could produce. T3 - The system correctly reviews the answers to open questions I enter. T4 - The system makes use of all the knowledge and information available to it to produce its advised grade to the input.

3. Perceived Understandability

U1 - I know what will happen when the system is used because I understand how it behaves. U2 - I understand how the system will assist teachers with decisions they have to make. U3 - Although I may not know exactly how the system works, I know how it will be used to make decisions about the grade. U4 - It is easy to understand what the system does.

4. Faith

F1 - I believe advice from the system even when I don't know for certain that it is correct. F2 - I have faith that the system will provide the best solution. F3 - When the system gives unusual advice I am confident that the advice is correct. F4 - Even if I have no reason to expect the system will be able to solve a difficult dilemma, I still feel certain that it will.

5. Personal Attachment

P1 - I would feel a sense of loss if the system was unavailable and would not be implemented. P2 - I feel a sense of attachment to the system. P3 - I find the system suitable for decision-making in open questions. P4 - I would like my educational institution to use the system for decision-making. P5 - I have a personal preference for a grading system with this system.