TRUST IN ARTIFICIAL INTELLIGENCE –
EFFECTS OF EXPLANATIONS AND
CLASSIFICATION ERRORS

VERTRAUEN IN KÜNSTLICHE INTELLIGENZ –
EFFEKTE VON ERKLÄRUNGEN UND
KLASSIFIKATIONSFEHLMERN

BACHELORARBEIT

im Studiengang Psychologie
vorgelegt an der
Fakultät für Humanwissenschaften
der Otto-Friedrich-Universität Bamberg

2018

Verfasser: Johannes Markus Burr

Gutachterin: Prof. Dr. Ute Schmid (Professur für Angewandte
Informatik insb. Kognitive Systeme, Fakultät
Wirtschaftsinformatik und Angewandte
Informatik)
Abstract

Artificial Intelligence has attracted broad attention since deep learning has overtaken the field of machine learning. Being black boxes, such systems face one critical issue: can the user trust the system? This thesis proposes that explanations play a vital role in the trust formation and, more importantly, for shaping appropriate trust. On different levels of automation, 132 participants interacted with a decision support system which was systematically varied in a 2 (explanation, no explanation) x 2 (error, no error) between-subject design. Both manipulations revealed a significant effect on Jian et al.’s trust scale (2000), which consisted of a two-factor structure. Moreover, reliance is examined based on a hybrid lens model (Seong & Bisantz, 2002). Evidence suggests that explanations enhance the performance of the user indicating more appropriate trust. Conclusively, the relevance to strive for comprehensible Machine Learning approaches is emphasized.

Zusammenfassung

Contents

1 INTRODUCTION ................................................................. 6
2 THEORETICAL AND EMPIRICAL BACKGROUND .............. 7
  2.1 Theoretical Framework ......................................................... 7
  2.2 Artificial Intelligence ............................................................ 9
    2.2.1 Levels of Automation ....................................................... 9
    2.2.2 Artificial Intelligence ...................................................... 11
    2.2.3 Machine Learning .......................................................... 11
    2.2.3.1 Neural Networks ......................................................... 12
    2.2.3.2 Decision Trees ........................................................... 13
    2.2.4 Why Explanations matter ................................................. 14
  2.3 Trust .................................................................................. 15
    2.3.1 Defining Trust ............................................................... 15
    2.3.2 Dimensions of Trust ....................................................... 17
    2.3.3 Emotional Facet of Trust ................................................ 19
    2.3.4 Operationalizing Trust .................................................... 19
  2.4 Learned Trust ................................................................. 20
    2.4.1 Explanation and Trust ..................................................... 21
    2.4.2 Reliability ........................................................................ 23
    2.4.2.1 Embedding diverse Findings ....................................... 23
    2.4.2.2 Signal Detection Theory ............................................. 24
    2.4.3 Can errors build trust? ..................................................... 25
  2.5 Situational Trust .............................................................. 27
    2.5.1 Appropriate Reliance ..................................................... 27
    2.5.2 A hybrid Lens Model ...................................................... 28
    2.5.3 Confidence as a Function of Trust ................................... 30
  2.6 Overview over the Study and Hypotheses ................. 31

3 METHOD ................................................................. 35
  3.1 Platform and Sample Size .................................................... 35
  3.2 Cover Story ......................................................................... 35
  3.3 Experimental Design .......................................................... 36
  3.4 Stimulus material .............................................................. 36
    3.4.1 Stimuli ........................................................................... 36
    3.4.2 Explanations ................................................................. 36
    3.4.3 System Errors ............................................................... 37
  3.5 Experimental Task ............................................................ 38
    3.5.1 Concept learning .......................................................... 38
    3.5.2 Sampling ....................................................................... 38
3.5.3 Error Feedback ................................................................................... 39
3.5.4 Decisions with Recommendation ................................................... 40
3.5.5 Transfer ............................................................................................ 40

3.6 Self-Reports .......................................................................................... 41
3.6.1 Trust ................................................................................................. 41
3.6.2 Confidence ........................................................................................ 41
3.6.3 Dimensions of Trust ......................................................................... 41
3.6.4 Enjoyment ......................................................................................... 42
3.6.5 Error Understanding ......................................................................... 42

3.7 Statistical Analysis .................................................................................. 42

4 RESULTS
4.1 Sample ..................................................................................................... 43
4.2 Reliability of Scales ............................................................................... 43
4.3 Tests on normal distribution ................................................................. 43
4.4 Trust Dimensions .................................................................................. 44
4.5 Trust ....................................................................................................... 45
4.6 Decisions with Recommendation ......................................................... 47
4.7 Transfer Task ........................................................................................ 49
4.8 Explorative Data Analysis .................................................................... 49

5 DISCUSSION
5.1 General Findings .................................................................................. 51
5.2 Separating the Trust Scale .................................................................... 51
5.3 Positive Trust and Distrust ................................................................... 52
5.4 Interaction on Trust and Confidence .................................................... 55
5.5 Trust Formation ...................................................................................... 56
5.6 Small Effect of Explanation .................................................................. 58
5.7 Adopted Trust model ............................................................................. 59
5.8 Appropriate Trust and Reliance ............................................................. 60
5.9 Limitations ............................................................................................. 61
5.10 Conclusion ............................................................................................. 62

6 REFERENCES ............................................................................................ 63

APPENDIX A: EXPERIMENTAL MATERIAL ........................................ 71

APPENDIX B: ADDITIONAL RESULTS .................................................. 75
1 Introduction

The rise of Artificial Intelligence is expected to entail profound changes to western societies as fast progress in Machine Learning facilitates practical implementations. A recent study found that at least one in ten jobs might be at risk of being replaced by automation (Dengler & Matthes, 2015). The prototypical future job might be supervising an artificially intelligent agent rather than mastering the task oneself. Highly automated systems are already in use in a variety of contexts. They manage enterprise’s logistics, have almost replaced human stock exchange speculators and are expected to drive our cars autonomously. Decision support system have gained impact especially in the field of medical diagnoses (Garg, Adhikari, & McDonald, 2005). What will change is that algorithms will be able to adjust and improve themselves and thus be applicable in more dynamic environments.

In Machine Learning, so called deep learning approaches are on the rise which face one major problem: they are purported black boxes. Being opaque, some individuals project supernatural powers into such systems while others reject them on principle. While there is a lot of research on improving accuracy, one major factor for real-life implementations is often overlooked: the human factor. No matter how well Artificial Intelligence performs technically - one major issue remains: Users will not use a system, which they do not trust (Ribeiro, Singh, & Guestrin 2016).

Trust is found to be a vital construct shaping the interaction between human and automation and hence must be considered in building automated systems. One crucial characteristic often ignored in constructing Artificial Intelligence, is the capacity to provide an explanation to increase the transparency of such systems. While some have addressed this as a necessity (e.g. Sinha and Swearingen, 2002), no comprehensive study about the effect of explanations on trust exists yet.

Besides a lack of trust, a second major problem in the interaction of humans with automation is the appropriateness of reliance (Parasuraman & Riley, 1997). Explanations might be especially valuable in imperfect systems when they help to detect incorrect recommendations.
2 Theoretical and Empirical Background

2.1 Theoretical Framework

First, the theoretical framework, in which this paper operates, will be presented. The subject of this thesis is the interaction between a human operator and an automated system, specifically an artificially intelligent agent (see Figure 1). The person pursues a goal determined by the situation, i.e. successfully handling a classification task, at which the system assists.

![General framework of human’s interaction with a decision support system](image)

Dzindolet et al. (2002) proposed the possibility to treat the “team” of system and user as a group in the social psychological sense. They assume both are more productive together than either one alone. This is the basic premise of this work, too. No task must generally be executed mechanically or manually, but instead the strengths of both should be merged as efficiently as possible. In fact, Nass et al. (1995) found that humans tend to convey their social schemata into the interaction with computers and hence treat them like human team members. This justifies that many researchers have adopted the construct of trust from interpersonal relations to the interaction between human operators and automated systems (Lee & Moray, 1992; Muir, 1994; Muir & Moray, 1996; Moray, Inagaki, & Itoh, 2000; Biros, Dali, & Gunsch, 2004; Parasuraman, Sheridan, & Wickens, 2008; Merritt et al., 2012; Hoffman, Johnson, & Bradshaw, 2013). Humans were found to use automation, when their trust in the automation exceeded their self-confidence to perform the task themselves (Lee & Moray 1994).

Trust can be further examined by focusing separately on the three components constituting the basic theoretical framework: the operator, the system, and the situation.
Respectively focusing on each one, three layers of trust proposed by Marsh and Dibben (2005) can be identified: human traits shape dispositional trust; features of the system generate learned trust; and context influences situational trust (Hoff & Bashir, 2015).

Dispositional trust refers to a trans-situational tendency differing between persons to put faith into other agents (Hoff & Bashir, 2015). Individual differences will be put aside in the present research. However, the distinction between learned trust and situational trust is relevant as the two main objectives of this study can be illustrated thereby. The differentiation between learned trust and situational trust is important in that there seems to be no obvious link between both (Lee and See, 2004). Other researchers have, likewise, distinguished between trusting an agent and trusting a single recommendation (Ribeiro et al., 2016; O’Donovan & Smyth, 2005).

**Learned trust** develops through experiences with the automated system and evaluations of those (Marsh & Dibben, 2005). Other researchers have coined the development of trust as trust formation (e.g. Lewandowsky, Mundy, & Tan, 2000; Lee & See, 2004). Generally, engineers should try to create as trustworthy systems as possible. Hancock et al. (2011) differentiate between attributes of the system, which were found to have an instable effect, and performance of the system, whose effect on trust was stable and large. The most researched antecedents of trust, belonging to performance-based characteristics, are predictability and dependability. Generally, when the behavior of the system is both desired and matches the expectation of the operator, trust is expected to grow (Muir, 1994). The first objective of this study is to investigate the influence of errors and explanations on the trust formation.

However, trust has an ambivalent facet as Cohen (2000) notes: “The problem of decision aid acceptance is neither undertrust nor overtrust, but inappropriate trust: a failure to understand or properly evaluate the conditions affecting good and bad aid performance” (p. 1). Humans are found to use automation as a social decision heuristic (Lee & See, 2004) and tend to overly rely on its recommendations which can lead to inferior performance compared to manual mode (e.g. Dzindolet et al. 2003). **Situational trust** refers to an adjustment of basic tendencies to trust due to situational cues (Marsh & Dibben, 2005). For real-life implementations, it is vital for humans to be able to judge the credibility of a recommendation in order to ensure a steady performance. The author of this thesis proposes to study reliance on a
micro level of single decisions, for which a hybrid lens model will be applied. It is alleged - as a second objective of this research - that explanations are especially weighty cues improving the performance of the human automation team.

### 2.2 Artificial Intelligence

This section focuses on the trustee, the Artificial Intelligence. Technical aspects are considered in so far as to distinguish between *white boxes* and *black boxes* as it has substantially different prerequisites for the trust formation.

#### 2.2.1 Levels of Automation

Today, machines execute a variety of functions once believed to be only manageable by humans and future developments are hardly foreseeable. *Automation* in its broadest sense refers to “machine execution of functions” (Parasuraman, Sheridan, & Wickens, 2000, p. 286). Some or whole parts of a task are transferred to an automated system superseding humans. This puts the person in a situation characterized by dependency and vulnerability in which trust can potentially arise.

HIGH 10. The system decides everything autonomously, ignoring the human

9. informs the human only if it decides to

8. informs the human if asked

7. executes automatically, then necessarily informs the human

6. allows the human a restricted time to veto before automatic execution

5. executes a suggestion if the human approves

4. offers a recommendation

3. narrows the selection down to a few

2. offers a complete set of decision/action alternatives

LOW 1. The system offers no assistance; human decides and acts

*Figure 2: Display of Levels of Automation of Decision and Action Selection adapted from Parasuraman et al. (2000)*

Sheridan and Verplank (1978) have proposed the first taxonomy of *levels of automation* (LOA). They define a continuum of machine autonomy, which they specify by describing ten distinct stages differing in function allocation (see Figure 2). The
first five levels refer to a manual mode with more tasks undertaken by the system upwards the scale. Levels six to ten represent an automated mode, where fewer information is provided to the human. As every level is characterized by a specific relationship between the system and the human, different levels of trust are expected. One might trust a system performing a task on level four but not on level eight. Generally, people tend to trust systems with a lower level of automation more (Lee & See, 2004).

Sheridan and Verplank’s framework focuses on decision making and execution. Parasumaran et al. (2000) have extended it by adding the stages information acquisition and analysis prior to the existing stages (see Figure 3). They propose examining a LOA scale for each stage. Most decision support systems are highly automated on the first three stages (a) information gathering, (b) information analysis and (c) decision making but leave (d) the implementation of the recommendation to the user.

Based on these four stages, Parasumaran et al. differentiate between information automation and decision automation. The former refers to an automatization up until the second stage. It provides the users with a compilation of information which is supposed to aid their decision making. In contrast, decision automation executes the task until the third stage and offers the users a definite recommendation for a decision without specifying the reasons. The authors speculate that the team of human and information automation is more robust to system faults because the operator is rather able to correct them. Decision automation would bear the danger of users “uncritically following incorrect advisories of imperfect automation” (Parasuraman & Wickens, 2008, p. 514). Black box systems can be assigned to decision automation as they solely provide recommendations. White box systems, which
will be further described beneath, are somewhat a combination of both. They provide a recommendation but are also transparent about the information they analyzed beforehand.

### 2.2.2 Artificial Intelligence

The field of research into *Artificial Intelligence* (AI) can be differentiated into two groups: ones that model human thought processes and ones that resemble human behavior (Russell, 2003). The first and older one, cognitive modelling, reflects upon computer simulations of human cognitive processes (Smith, Kosslyn, & Barsalou, 2008). It has several interesting theoretical implications for psychology which will not be discussed in this work. Instead, the second field, the engineering perspective of AI, will be further examined. These algorithms are inspired by human cognition but not constrained to it. The goal is to build systems that can solve tasks which were initially thought to be only manageable by humans (Rich & Knight, 1991). A famous operationalization stems from Alan Turing (1950) who proposed to call a machine intelligent when it is indistinguishable from an undeniably intelligent agent – a human. To utterly pass this test, the machine would not only need to be able to represent knowledge, draw new conclusions and communicate, but also to adapt to dynamic changes in the environment, i.e. learn (Russell, 2003).

### 2.2.3 Machine Learning

Mitchell (1997) describes the field of *Machine Learning* to be “concerned with the question of how to construct computer programs that automatically improve with experience”. He suggests Machine Learning’s scope in (a) data mining (b) insufficiently understood domains and (c) dynamic contexts.

Similarly, Michie (1988) defines Machine Learning as “generating an updated basis based on sample data for better predicting subsequent data”. However, this is only the weak criterion he proposed. He suggests the evaluation of Machine Learning on two orthogonal axes: accuracy and comprehensibility. The strong criterion adds the ability to “communicate its internal updates in explicit symbolic form”. If the user additionally understands this communication the ultra-strong criterion is met. (as cited in Schmid et al., 2017). As accuracy could be measured more easily, the question of comprehensibility was mostly neglected. However, today, as practical
implementations are imminent, the relevance of understanding the algorithm is gaining more interest. Most researchers agree, that accuracy and interpretability contribute a trade-off (Alcalá et al., 2005; Kamwa, Samantaray, & Joos, 2012). Deep learning algorithms usually reach a better fit but are incomprehensible, whereas decision trees, for example, are easily interpretable but tend to be less accurate (see Figure 4).

“Black box” refers to an opaque system, where only input-output processes can be observed. The inner workings are not accessible. A white box, in contrast, is transparent, and different steps of processing can be monitored and potentially understood. In the following, technical examples for both will be described. The inherent difference lies in the fact that black boxes are not comprehensible, whereas white box systems are. Schmid et al. (2017) recently proposed an operational definition, which will be adopted for the present research. They define the comprehensibility of a specific program P to be the “mean accuracy with which a human […] after brief study and without further sight can use [the program] to classify new material sampled randomly from the [program’s] domain.” (p. 3).

2.2.3.1 Neural Networks

Artificial neural networks (ANN) were originally inspired by the structure of neurons in the human brain (Russel, 2003). Mitchell (1997) remarks its scope of application in the processing of noisy sensor data, where long training times are acceptable, and it is bearable for humans not to understand the target function. It is
especially useful in tasks which humans mainly process implicitly, like face recognition, and thus cannot be engineered based on hand-picked features.

The smallest element of an ANN is a perceptron which has several vectors or real-valued inputs from which it calculates a linear function. It has a dichotomic output discerned by a specific threshold. However, a linear function cannot discriminate every kind of data successfully (Mitchell, p. 87).

Multiple perceptrons can be combined to a multilayer network which extends the system’s computations by several dimensions. Additionally, neurons usually compute sigmoid functions generating non-linear relationships. Further output-neurons and so-called hidden layers, which are neither input nor output neurons, can be added. They offer the advantage of inventing new features that are not explicit in the input data (LeCun, Bengio, Hinton, 2015). In the 1970 to 1980 several groups independently discovered that multilayer architectures can be trained by simple stochastic gradient descent. The so-called backpropagation procedure is an application of the chain rule for derivates. Gradients are computed beginning at the output layer, to subsequent modules back to the input layer. This way, the model calculates a best fit. A possible difficulty can be local minima in which the target function could be stuck. However, recent theoretical and empirical findings suggest this problem is slighter in large networks than previously thought (LeCun et al., 2015).

Another relevant distinction is between supervised and unsupervised machine learning. The former requires pre-labeled data with which the system is fed, so that an error function can be calculated. It has overshadowed unsupervised learning in recent years. However, unsupervised learning is expected to gain relevance in the near future as it is closer to dynamic human learning and more efficient in the industrial context (LeCun et al., 2015).

2.2.3.2 Decision Trees

The AI system assisting participants in the present study can be modeled by a Decision Tree. It classifies stimuli into two groups based on three characteristics (see Figure 5). Its background will be presented as an example for white box learning.
Decision Trees are one of the most popular inductive inference algorithms. They avoid problems of restricted hypothesis spaces by searching a completely expressive hypothesis space. The ID3 algorithm, for example, constructs its decision trees top-down (Qinlan, 1986).

First, it identifies the mightiest attribute by evaluating how many training examples can be correctly classified by each attribute alone. It becomes the root node of the tree. A descendant is then created for both possible values of the attribute above and the training examples are separated accordingly. Then, the same statistical test to find out the best single classifier for the remaining groups is identified. This results in an algorithm which never backtracks (Mitchell, 1997). One major advantage is the possibility of transforming its discrete-valued target function into if-then rules, which improves human readability and thus comprehensibility (Mitchell, 1997). This way, logic approaches can communicate their inner status, which offers the edge of offering an *explanation*.

### 2.2.4 Why Explanations matter

One might argue that it is irrelevant to demand to understand a system which performs flawlessly. Yet, numerous concerns arise. If one blindly trusts in the accuracy of a system, both ethical and practical issues can occur.

Ribeiro, Singh, and Guestrin (2016) argue that explanations help to assess the capacity of new systems better than simply examining their accuracy scores on validation sets. Here, researchers would usually overestimate the real accuracy, for
example due to data leakage. Explanations might help to assess the basis, the theoretical ground, on which the system operates.

Second, if it is unclear – even for the designer – how a system operates, it is possible that it reaches decisions based on undesired biases. If input data includes racist biases – for example one ethnic group being more prone to crime than another – the system will adopt and implement this bias independent from its cause. As it cannot communicate its inner processes this bias is neither detectable nor adjustable; the system lacks insight. This is why the EU has passed the General Data Protection Regulation (GDPR), which secures the right to obtain a meaningful explanation of the logic involved” when automated decision making takes place (Guidotti et al., 2018).

Third, while a black box can help to improve performance, it cannot drive progress. The idea to only focus on accurate predictions resembles the scientific idea of instrumentalism. Some scientists believe the mere purpose of science is to predict future events. Whereas predictions undoubtedly are a cornerstone of science it is not its ultimate goal. Instrumentalists reject the necessity of functional explanations as superfluous, which is incorrect. Imagine a Deep Learning model built to predict physical events. Let it be so good it predicts all physical phenomena perfectly. This model would become a simulation of reality, which would undoubtedly be impressive. Anyway, what would it help? It would replace laboratories but not experiments. It could neither promote theoretical advantages nor practical applications, because for both theoretical considerations are still vital. While this model would contain all the answers, we would still lack the questions that need to be asked. What it lacks is understanding (Pearl & Mackenzie, 2018).

2.3 Trust

After contrasting white boxes to black boxes, now trust as a theoretical construct is defined and operationalized.

2.3.1 Defining Trust

While there is no consensus in how to define trust, most researchers agree that it is a multidimensional construct (Barber, 1983; Rempel et al., 1985; Seong, Bisantz, 2002; Lee & See, 2004). Most definitions are adopted from the literature on
interpersonal trust, which originally developed from sociological literature (Llinas et al., 1998). One of the broadest, yet condensing, definitions stems from Deutsch (1973) who stated that trust is “the confidence that one will find what is desired from another, rather than feared” (p.148) (as cited in Muir, 1994). This emphasizes a core requirement: the situation needs to be characterized by some uncertainty. Luhmann (1978) proposed the function of trust to be a reduction of uncertainty. Without trust, one would have to consider every possible – even the most implausible – threats in the social interaction with other persons. Trust eliminates some of those options enabling to consider the remaining ones more closely. Nass, Fogg and Moon (1996) found that interpersonal and human-automation trust are in some sense similar. One explanation is that the second is just a special form of the first. In some way, if the user trusts a system, he trusts the designer of the system who is just one step further away (Hoff & Bashir, 2015).

Castelfranchi and Falcone (2010) note that the plurality of definitions can be distinguished into two groups: (i) those that focus on trust as a mental state concerning an attitude of a person towards another person or surrogate and (ii) those that focus on the behavior caused by this attitude as relying on the opposite. Most definitions from (i) focus on expectations. For instance, Rotter (1967) defines trust as the “expectancy held by an individual that the word, promise or written communication of another can be relied upon” (p. 651). Similarly, Rempel et al. (1985) describe trust as the “generalized expectation related to subjective probability an individual assigns to the occurrence of some set of future events” (p. 96). However, a definition merely focusing on expectations lacks the dimension of competence. Such a definition may make the concept of trust superfluous as an agent steadily behaving equally, ignorantly to the context, may be highly predictable but not useful nor trustworthy (Castefranchi & Falcone, 2010).

The second group of definitions focusing on (ii) the behavior often characterize trust as a willingness to rely on an opposite accepting some vulnerability (Lee & See, 2004). For instance, Moorman, Deshpande, and Zaltman (1993) refer to trust as a “willingness to rely on an exchange partner in whom one has confidence” (p. 82). Some studies operationalize trust purely by measuring reliance, i.e. following a recommendation (e.g. Ribeiro et al., 2006). The problem here is that reliance is no immediate result of trust (see 2.5.1.).
One definition covering both components – (i) the mental state and (ii) the behavior was contributed by Madsen and Gregor (2000): “Human-computer trust is defined […] to be 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 agent.” (p. 1). Such a definition, however, without explaining the functional link between the mental state and the behavior is fragmentary. One exemplification offers the model of Ajzen and Fishbein (1980), which allows keeping beliefs, attitudes, intentions, and behavior conceptually distinct. Nevertheless, it makes no precise predictions. This is why the present thesis separates the mental state and the behavior. Trust will be narrowly defined as a mental state and the link towards the behavior will be examined separately focusing on reliance.

One of the most accepted definitions stems from Lee and See (2004) who define trust as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (p. 54). This definition was chosen for the present research, because it is condensing multiple relevant aspects into little expression. It establishes the operators to follow a specific goal for which they need to decide under uncertainty. Trust is defined as an attitude which includes an evaluation of the helpfulness of an agent. Note that the described attitude does not demand relying on the recommendation. Neither is it required that the recommendation must be correct. Lee and See (2004) add that this definition needs to be elaborated in terms of the appropriateness of trust, which they specify in terms of calibration, resolution and specificity, the cognitive processes that govern the trust formation (see 2.4.), and the influence of context (see 2.5.).

2.3.2 Dimensions of Trust

Lee and Moray (1992) identified three general dimensions of trust: performance, process and purpose. They did an extensive literature review convincingly integrating a plurality of models from different researchers into these three components, which are well accepted within the literature.

*Performance* is defined as the “expectation of consistent, stable and desirable performance or behavior” (Lee & Moray, 1992; p. 1246). It generally explains what the system does and is based on a competence demonstrated by successful task
processing. It is linked to characteristics like predictability and reliability (Lee & See, 2004).

Process which “depends on an understanding of the underlying qualities or characteristics that govern behavior” (Lee & Moray; p. 1246) refers to an understanding how the system operates. If the algorithm can be understood and appears appropriate for a successful handling of the task, trust can be based upon (Lee & See).

Purpose reflects the “designer’s intention in creating the system” (Lee & Moray; p. 1256). Why a system was designed gives some assurance of its positive orientation towards the trustor and thus can establish faith (Lee & See).

However, the three dimensions are not clearly separated; interactions between them are possible. For example, conclusions about the inner workings of a system can be derived by observing how it performs. Accordingly, inferences about the intentions of the designer can be drawn. Likewise, this sequence could take place in reverse. Knowing about the purpose of a system allows forming expectations about the internal processes which can be validated by monitoring the behavior (Lee & See, 2004). This model suggests that it is possible to build trust into a black box system when one experiences it to constantly perform accurately or when one is being assured of positive intentions of the designer. These two points of contact are usually the only ways to establish trust. However, when a system is transparent, users can observe its inner workings, from which trust might arise likewise.

Consequently, trust can be based on any of these three dimensions or on several ones simultaneously. McKnight, Cummings, & Chervany (1998) tried to resolve conflicting findings that either described trust as fragile or robust. They argue that trust is likely to be robust when several of its antecedents are high. In contrast, when trust is based on few or one factor it can be instable. Even though the antecedents they proposed are not identical to Lee and Moray’s (1992) dimensions of trust they can be easily integrated. Trust was found to be especially robust when it is based on an understanding of fundamental motives of the agent in contrast to when it rests merely upon its reliability (Rempel et al., 1985). This implies interacting effects between the dimensions of trust.
2.3.3 Emotional Facet of Trust
In the context of organizational science, trust is considered from a rational choice perspective (Hardin, 1992). In the classical Prisoners Dilemma trust would be operationalized as the decision to cooperate, which is traditionally being coined as irrational (Robinson, 1975). In contrast, Lee and See (2004) denote trust as a primarily affective response, also influenced by analytical and analogical processes. They emphasize the importance of emotions in decision making and implicit processes of communication. They derive that trust is formed and functions mainly through emotional processes. Ultimately, they conclude that the display connecting both interaction partners is vital for trust to arise. Display features, their organization, and image quality might play a key role influencing the perception of trustworthiness.

2.3.4 Operationalizing Trust
As there are multiple definitions of trust, researchers chose different operationalizations. The very first experiments by Lee and Moray (1994) used single-items measurements asking participants “How much do you trust the system [performing a specific action]?”. However, for single-item measures, no Cronbach’s Alpha can be calculated, and they are considered deficient (Wanous, Reichers, Hudy, 1997). Others tried to measure trust through behavioral measures. For instance, a higher response-frequency (Spain & Bliss, 2008), reliance, i.e. following a recommendation (Ribeiro et al., 2016), or choosing a higher level of automation (Lee & Moray, 1994), were interpreted as trust behavior. In contrast, the present study defines trust as a mental attitude and thus it is measurable through self-report. Many researchers have used self-report (Atoyan, Duquet, Robert, 2006; Madhavan & Wiegman, 2007; Desai et al., 2012; Merritt et al., 2012), yet, no generally accepted questionnaire exists.

One of the most used questionnaires of trust in automation is the one by Jian, Bisantz, and Drury (2000)1. The questionnaire was developed exploratively through a lexical approach. This fits to the origin of the concept of trust from everyday language. Thus, no theoretical assumptions by researchers come along like in other, rather normative approaches (e.g. Madsen & Gregor, 2000). Jian et al.’s

---

1 Cited 524 times according to google.scholar on the 21.08.2018.
questionnaire focuses on a purely subjective evaluation of the helpfulness of an agent by items like “The system provides security” or “The system has integrity” or negatively framed questions like “The actions of the system have harmful or injurious outcomes”. The questionnaire contains seven items measuring trust and five measuring distrust, all specifically designed for human-automation interaction. The authors argue that the high negative correlation between both indicate that trust and distrust are located on one dimension representing one construct rather than a two-factor model. Other researchers showed good reliability and validity (Safar & Turner, 2005). In the following, the three-step empirical procedure by Jian and colleagues to develop the questionnaire will be described.

First, in a word elicitation study, seven graduate students in Linguistics or English wrote descriptions of their understanding of trust and distrust and rated 138 words in how far they were related to trust. This resulted in 112 words that were used in the second phase, the questionnaire study. 120 students rated the extent to which those words were related to trust or distrust. The results indicated that trust and distrust were highly negatively correlated as ratings for each word correlated with $r = -.95$ in the human-automation conditions. The authors interpreted this as evidence that both constructs can be described to lie on one dimension. During the first and second phase they, additionally, differentiated between trust and distrust in general, which showed that trust in general and human-machine trust do not differ significantly while the other pairs did. The results of phase 2 provided 15 words related to distrust and 15 related to trust. These were used in phase three in a paired comparison study. Thirty students rated all word-pairs on a seven-point Likert scale ranging from “Totally different” to “Almost the same”. The results were analyzed in a factor and cluster analysis which produced a 12-item scale with two subscales for trust and distrust. The authors argue that the subscales indicate a single underlying factor lying on a continuum.

### 2.4 Learned Trust

Based on Lee and Moray’s (1992) three dimensions of trust - performance, process, and purpose - the trust formation is discussed (see Figure 6). Theoretical
considerations and evidence to the effects of explanations and errors on trust are discussed, after which a possible interaction is proposed.

2.4.1 Explanation and Trust

Transparency refers to the circumstance when “the inner workings or logic [used by] the automated systems are known to human operators to assist their understanding about the system” (Seong & Bisantz, 2008, p. 611).

One possibility to facilitate transparency is a constant feedback of task performance (Dadashi, Stedmon, & Pridmore, 2012; Gao & Lee, 2006). Having said this, it is impractical to be implemented in practice since decision-support systems are operated under uncertainty and without immediate feedback. Another possibility is to report an estimation of the probability how likely the recommendation is correct (Sinha & Swearingen, 2002). Lastly, an explanation can be added to the recommendation. Note that only logical approaches of Machine Learning (see 2.2.3.2) are capable of providing such an explanation.

There are different forms of explanations. Sørmo et al. (2005) differentiate five objectives of explanations in AI:

- Justification (Why?)
- Transparency (How?)
- Relevance (Why this question?)
- Conceptualization (clarify meaning)
- Learning (teaching).

![Figure 6: General overview of the effects of explanations and errors](image)
Wang and Benbasat (2007) showed that different types of explanations affect different dimensions of trust. Pieters (2010) proposes justification or “offering reasons for an action” (p. 60) to be the main goal in expert systems. This will be implemented in the present research through the explanations further elaborated in section 3.4.2. They emphasize the relevant characteristics of the stimuli which determine its category membership in everyday language. However, such an explanation not only explains the why but also conceptualizes the categories and thereby enables the user to learn.

Glass, McGuinness, and Wolverton (2008) inquired into basic concerns of users regarding technology with qualitative methods. They concluded that most of the trust issues the participants identified could be solved by providing an explanation. Similarly, almost all participants (86%) in an experiment of Herlocker, Konstan, and Riedle (2000) wanted explanations to be constantly included in a recommender system. Others have also addressed the necessity of explanations in decision support systems (e.g. Sinha and Swearingen, 2002; Pieter, 2011), yet it is still undervalued in engineering AI and no comprehensive study on the effect on trust exists, yet.

Explanations creates transparency and thus into the first and second stage of task fulfillment proposed by Parasumaran et al. (2000) (see 2.2.1.). Thereby, it is more understandable, how the system operates. This way, the dimension of process is expected to be affected (see 2.3.3). Explanations should help to convince the user, that the system functions in an effective way. Furthermore, derivations about future performance and the purpose of the system should be facilitated. Overall, trust is expected to grow.

One of the few experimental findings on this topic stems from Dzindolet et al. (2003). He systematically varied his participant’s understanding of system errors. Whereas all were informed that the system was not perfect, half of the participants were provided with a rationale why errors might occur. They were told that the system was a pattern detector searching for human-like structures. When the shadows or trees had human-like structure, it might confuse them. It resembles an explanation, which certainly enhanced the transparency of the system. In fact, those who received the rationale reported higher trust and relied more on the automation. However, in contrast to the present experiment, this rationale acts on a higher level
since it was only provided once. In the present study, on the other hand, the effect of explanations, which are provided steadily, will be inspected.

2.4.2 Reliability

2.4.2.1 Embetting diverse Findings

A lower reliability is generally believed to minor trust. Errors directly influence the helpfulness of the opposite, and if they are correctly perceived, it should directly influence the assessment of the trustworthiness. However, the perception of such errors seems to affect their influence, too. Moreover, as noted before, denoting trust purely as an assessment of a performance is insufficient and ignores the dimensions of process and purpose.

Lee and See (2004) propose a dynamic model in which the impact of an error declines exponentially with time. Trust would drop abruptly after an error occurs but could recover with time. This model focuses on single errors and less on the general effect of defective performance. They also note that most models, instead, have a static approach. As literature lacks a comprehensive model, most researchers have adopted general cognitive theories to apprehend their findings. Findings suggest a more complex relationship than an obvious linear link between errors and trust.

For instance, the timing of an error influences its impact. Errors which arise early in the interaction have bigger impact than ones later in the interaction (Manzey, Reichenbach, & Onnasch, 2012). This implies that trust calibration is rather stable after a first impression has been made. Similarly, Llinas et al. (1998) describe “two vicious cycles on the trust continuum” (see Figure 7). They argue that overtrust stabilizes itself because relying on the automation steadily leads to a degradation of skill and confidence, which reinforces overtrust again. Likewise, distrust leads to a

![Figure 7: Two vicious cycles on the trust continuum; adapted from Llinas et al. (1998)](image-url)
manual task performance, which prevents evaluating the system, and thus an update of the primary evaluation is unlikely. Secondly, when a user perceives the task as easy in which the system errs it will have more impact than in tasks perceived as difficult (Madhavan, Wiegmann & Lacson, 2006). This finding can be attributed to humans transferring their social schemata into the interaction with automation (Nass, Fogg, & Moon, 1996). They might expect automation to work like human processing and therefore simple errors indicate a generally flawed system (Hoff & Bashir, 2015). Thirdly, Vries, Midden, and Bouwhuis (2003) manipulated errors both in a manual and automated mode based on the assumption of Lee and Moray (1994) that trust and self-confidence are antagonists. They found that errors had stronger effect on trust than on self-confidence. They explain this effect with the generally shared expectation of automation to be near perfect. The underlying schema makes mistakes more salient for the system while personal errors are more easily integrated or displaced.

2.4.2.2 Signal Detection Theory

One theory, to which researchers usually refer to is the classical signal detection theory. It classifies alarms based on an objective reality. Assuming the system has two different states, alarm and no alarm, the correctness of those states are determined by the context they arise in which is either adequate or inadequate. This results in a 2 x 2 matrix with the classical differentiation between Hit, False Alarm, Correct Rejection, and Miss (Swets & Green, 1966).

The meta-analysis of Schaefer et al. (2014) found no difference between False Alarms and Misses. In contrast, Hoff & Bashir (2015) suggest different effects on trust. They argue that false alarms being more salient should have more influence on trust, which is supported by some (Johnson et al., 2004), yet disputed by other research (Madhavan, Wiegman, & Lacson, 2006; Rovira & Parasuraman, 2010). However, Hoff & Bashir (2015) also conclude that the 12 reviewed studies did not reveal a clear difference in the effect of False Alarms and misses. Instead, the “consequences of each error type in a given environment” (p. 426) seem to be most significant. They propose future research “should [...] consider studying how operators naturally react to preplanned automation breakdowns” and “systematically manipulate the negative consequences” (p. 428) of errors in order to test the hypothesis that the consequences are crucial.
Salem et al. (2015) offer supportive findings from the field of human-robot interaction. Participants interacted with a robot in four different situations in which the robot either performed these tasks correctly or faulty. For instance, after greeting the persons the robot proposed to show them the way to the sofa. It either took a straight path or a detour and abruptly spun around itself in the faulty condition. While Salem and colleagues found differences in subjective ratings, participants did not differ significantly in how they behaved. The robot requested them to perform five unusual acts. Surprisingly to them, participants in both conditions equally complied, which Salem and colleagues interpret as an objective measure of trust (see 2.3.5.) However, there was a significant effect for task type. For illustration, only 10% did not comply with a request to throw letters into a bin beside the table, whereas 32.5% refused to pour orange juice over a plant. Salem and colleagues interpret the tasks to differ in their harmfulness. Placing letters in a bin is revocable as opposed to spilling orange juice, which supports the claim that the severity of errors determines their impact.

2.4.3 Can errors build trust?

Lees and Lee (2007) argue that the distinction based on the signal detection theory is not sufficient. Instead, they propose a framework focusing on the perspective of both the designer and the user. It is based on the three dimensions of trust from Lee and Moray (1992): performance, process and purpose. Lees and Lee manipulated two of them in a driving task where participants were assisted by a collision warning system. Participants faced a system with differing error rates (performance) which occurred in different contexts (process). They argue that False Alarms (FAs) reflect a break down on both dimensions while unnecessary alarms (UAs) only on the first since they provide a chance to understand how the automation works. They explain UAs in the following: “An alarm associated with a situation judged hazardous by the designer, but not by the driver. The driver can understand what triggered the alert.” (p. 1267) Both would be assigned to False Alarms in Alarm Detection Theory. They found that FAs diminished trust while UAs even increased trust slightly without meeting standard criteria for significance. Apparently, UAs did not diminish trust because they provided insight into how the automation operates and thereby enhanced understanding.
This implies that errors might even increase trust, when they do no harm, but instead offer an opportunity to learn, how the opposite behaves. In Lees and Lee’s study, participants did not depend on the system in such situations, making the alarm superfluous. When a situation is not uncertain, trust cannot arise in the first place. Similarly, explanations are expected to increase participants knowledge about the task, which then again should enable them to handle the task on their own. Lee and See state that “trust can develop when a systematic fault occurs for which a control strategy can be developed.” (p. 72). They note that “the effect of automation fault depends as much on its predictability as on its magnitude”. They underpin this claim with evidence that a constant error with a large fault had less impact on trust than a small error with unpredictable results (Moray et al., 2000; Muir & Moray, 1996). An anecdotal example about a printer in the University Bamberg’s library might explain this idea intelligibly. For some time, many students complained it regularly crashed, when they wanted to print. Today, this problem is solved with a sign reading “If you want to print more than 20 pages, please use the other printer. It will collapse otherwise”. With this easy control strategy, the printer does not appear to be defective anymore, since it perfectly carries out the task of printing pages under the amount of 20. Knowing this, students can easily trust it for this specific task.

What is different in a classification task to Lees and Lee’s study, is that an error is never intended by the designer. Instead, “Fortuitous Alarm” from Lees and Lee’s framework refers to an unintended, yet predictable and useful error. The authors also suggest, that fortuitous alarms might increase both performance and trust. Recall that having defined trust as an attitude towards the helpfulness of another, does not demand the recommendations to be objectively correct and neither does it demand the user to rely on them. It seems plausible, that an incorrect recommendation might still carry the information necessary to the operator to solve the task successfully. For this, the user would need to understand the system’s errors, so that they are predictable and useful. Then, a Fortuitous Alarm would occur, which is expected to enhance trust. Together with evidence mentioned before, these considerations imply an interaction between errors and explanation. When participants are capable of understanding the errors, they should be both able to predict and counterbalance them, and thus decrease their severity.
2 Theoretical and Empirical Background

2.5 Situational Trust

After defining Trust and describing how trust is formed (learned trust), now the appropriateness of reliance under the concept of situational trust is discussed.

2.5.1 Appropriate Reliance

While the increase in automation offers many opportunities, it also opens new risks. Rarely do humans rely on automation perfectly. Instead, humans tend to disuse and misuse automation. Disuse refers to the failure of rejecting a system despite its capabilities. It is linked to a lack of trust in the system and thus underutilizing it. Misuse, on the other hand, occurs when the operator relies on automation inappropriately (Parasumaren & Riley, 1997). Overreliance is so common, researchers have coined the term automation bias which refers to “the tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing” (Mosier & Skitka, 1996, p. 205). Relying on automation seems to act as a social decision heuristic overcoming cognitive complexity (Lee & See, 2004; c.f. Dzindolet et al. (2003) for other social psychological explanations).

In their comprehensive review Lee and See (2004) point out that supporting appropriate trust is vital in avoiding misuse and disuse. However, they also note that considering reliance as a binary process is merely a simplification that makes its discussion more tractable. Researchers initially suspected a linear link between trust and reliance (Wiegmann, Rich, & Zhang, 2001). However, the relationship was found to be more complex. Other factors, like self-confidence, (Lee & Moray, 1994), workload (Riley, Lyall & Wiener, 1994) and environmental constraints influence the link. For example, one might be obliged to act according to a recommendation without trusting its source, which could be denoted as delegation (Castelfranchi & Falcone, 2010). On the other hand, one might trust a source but a lack of time or an imbalance between costs and benefits prevents considering the recommendation (Lee & See, 2004).

Llinas et al. (1998) propose to operationalize the appropriateness of trust through performance. They refer to the accordance between the level of trust by the operator and the actual trustworthiness as calibration of trust. However, they note that such an accordance would be difficult to assess as no measure for objective trustworthiness exists. Instead, they propose a performance-oriented perspective. They argue
that from an engineering standpoint, optimizing the performance of the interaction between human and automation has always been the objective of trust research. The better one performs with the help of a system; the more appropriate one’s assessment of the system’s trustworthiness seems to be. Thus, a better performance can be interpreted as more appropriate trust.

Lee and See (2004) conclude that appropriate trust “depends on the operator’s understanding of how the context affects the capability of the automation”. Reliance strategies can never be optimal per se. Like any behavior, it needs to be adopted and adjusted to particularities of the situation. This is why, situational trust is referred to here. Reliance represents a single decision, which massively depends upon the situation. Accordingly, Lee and See (2004) highlight the influence of context. The situation both influences the capability of the system (as it determines the task) and the human (e.g. mental workload) and thus has a moderating influence of trust on reliance. For further investigating the influence of situational cues, the lens model will be applied.

2.5.2 A hybrid Lens Model

The lens model is used in multiple lines of research to model human judgement. Likewise, it can also be transferred to the situation when operators decide whether to rely on an AI decision support systems’s recommendation.

Generally, the lens model separates the objective state of the world from so called cues. As the true state of the world is not accessible for humans, they make their judgements based on observable cues. They are, in turn, to some degree associated with the objective state of the world. In this way, judgmental errors are explainable by a lack of validity of utilized cues (Brunswik, 1955).

Seong and Bisantz (2002) first implemented the lens model into the context of human-automation trust. They propose a hybrid model combining a n-system design and a hierarchical system design (see Figure 8). On the one hand, the users inspect the stimuli and can derive their own judgement, which represents the usual n-system design. What is special, is that, on the other hand, the operators stand one step
behind the system which, too, processes the task. The system’s recommendation becomes a cue itself for the human’s judgement.

The hybrid lens model combines both paths and illustrates why it is so hard to study reliance. It is never entirely clear based on which cues the user decides. Even when the human takes the same decisions the system has recommended, this might simply be due to her deriving the decision on her own using first level cues. Yet, in any known paradigm, it would be denoted as reliance.

Automation bias can be specified as an excessive usage of automation cues. First level cues might be neglected or at least have less influence on the decision. It is plausible that the recommendation acts as an anchor after which only few additional cues are inspected for adjustments (Tversky & Kahnemann, 1974). Having defined the human-automation team as a social group suggests that a recommendation might even alter the perception of the first-level stimuli similar to what is found in human groups (Sherif, 1937).

The benefit of the hybrid model is its clarification of the advantage of explanations: explanations connect first and second level cues. The validity of the recommendation can be examined by matching its content with first level cues. When an explanation states that a stimulus belongs to category B because it is round, but the user

![Figure 8: A hybrid Lens Model representing the decision of a human operator; adopted from Seong & Bisantz (2002) (Image)](image)
observes a narrow form, the recommendation should arguably appear less trustworthy. Therefore, wrong recommendations should be identified more easily, which is much harder in opaque systems. This way, explanation should enhance the performance of the human operator, which would indicate more appropriate trust.

This is supported by Parasuraman et al. (2000) discussion comparing information automation and decision automation (see 2.2.1). False advice seems to have more impact, when it is provided as a recommendation than as information (Rovira, McGarry, & Parasuraman, 2007). For instance, Sarter and Schroeder (2001) tested pilots using an icing-monitoring automation, which either provided a decision recommendation or a status information. When reliable, both systems enhanced performance. However, false advice had higher impact, when a recommendation was provided than when there was status information.

Ribeiro et al.’s (2016) study based on their LIME algorithm offers support for this claim. They trained a logistic regression model with hand-picked stimuli, so that it classified between dog and husky based on whether snow was in the background. Participants were presented ten pictures and the decision of the model, where two errors occurred. They were asked whether they trust the model to work well in the real world and what they think how the algorithm operates. In fact, after the explanation was added almost all participants understood how the algorithm worked, which lead to a steep drop of trust.

### 2.5.3 Confidence as a Function of Trust

Trust and Confidence are sometimes viewed as synonyms in the HCI-literature (Schaefer et al, 2014). However, other researchers differentiate both constructs on numerous dimensions. For the present research the distinction by Siegrist, Earle and Gutscher (2003) is adequate. In the context of risk management, they propose a dual-mode model of social trust and confidence with the goal of predicting cooperation. They define cooperation vaguely, which can be translated to reliance. They define trust to arise in a relationship between two agents when one accepts some vulnerability based on “a judgement of similarity of intentions or values” (p. 706). Thus, trust is shaped emotionally whereas confidence is mainly a cognitive result.

*Confidence* arises between an agent and an object. It is a cognitive assessment of the probability of future events. Griffin and Tversky (1992) define confidence as a
degree of belief in a hypothesis based on evidence. Here, confidence in single
events not in general performance or knowledge is meant.

Defining the reduction of uncertainty to be the function of trust, confidence can be
regarded as the result. Confidence in decisions with recommendation might either
stem from own knowledge, or trust in the system. Likewise, Siegrist et al. (2003)
added a path from trust to confidence in their statistical model based on the data.
They speculate that the emotional component influences the perception of risk.

2.6 Overview over the Study and Hypotheses

The experiment ought to shed light on the interaction of humans and an AI decision
support system in a classification task. It was important to find a practical scope in
which the participants had equal – or no - prior knowledge and were generally in-
favor to the systems so that trust could arise. Stimulus material differed dichoto-
mously on four dimensions forming fuzzy stimuli that could still be clearly classi-
fied into two categories.

Participants learned the structure of the categories through the classical 5-4 presen-
tation by Medin and Schaffer (1978). Then they handled a classification task on
three different levels of automation (see 2.2.1). First, they observed the system clas-
sifying nine stimuli on its own (LOA 10) constituting a sampling, which was fol-
lowed by an error feedback. Then, participants classified seven items themselves
being provided with a recommendation by the system (LOA 4), after which they
decided about seven stimuli on their own (LOA 1). The capabilities of the system
participants faced were systematically varied in a 2 (Transparency: explanation, no
explanation) x 2 (Reliability: 31.25% errors, no errors) between-subject design.

The first objective of this study is to investigate the effects of the manipulations on
the trust formation. The first hypotheses are relatively straight forward: errors minor
the helpfulness of an opposite and thus directly influence the assessment of trust-
worthiness; explanations, on the other hand, assure the operator of a functioning
algorithm. (see 2.4.1).

\[ H1 \text{ Participants whose system provided an explanation report higher trust.} \]

\[ H2 \text{ Participants whose system was faulty report less trust.} \]
Trust is measured through self-report after participants classified stimuli with the aid of the system. Jian et al.’s trust scale (2000) is used. The authors claim that both subscales, positive Trust and Distrust, constitute two ends of a continuum.

\[ H3 \text{ The subscales of the trust scale are highly negatively correlated.} \]

Additionally, the dimensions of trust theorized by Lee and Moray (1992) are measured as manipulation checks. The differing reliability is expected to affect the dimension of performance, whereas the manipulation of explanations is supposed to affect the dimension of process. As the purpose of the system was not actively varied, no effects between the groups are expected.

\[ H4 \text{ Participants in the white box group report the system to be more understandable.} \]
\[ H5 \text{ Participants in the error group report the system to be less reliable.} \]
\[ H6 \text{ Participants in all groups report equal faith in the system.} \]

Besides, predicting trust through the subjective measures of its three dimensions is explored.

\[ H7 \text{ Measurements of perceived understandability, faith, and perceived reliability are linked to trust.} \]

Furthermore, Lee and See’s (2004) hypothesis that trust mainly consists of an affective component is tested. For this, participants report how much they enjoyed the classification task.

\[ H8 \text{ The trust scale and the measure of perceived enjoyment are positively correlated.} \]

Furthermore, an interaction on trust between the variation of reliability and transparency is expected. Several theoretical considerations imply that errors have less impact on the trust formation when explanations are being provided. First, trust was found to be more stable, when it is based on transparency (Rempel et al., 1985). Second, the severity of errors seems to determine their impact on trust (Lee & See, 2004). An explanation should help to detect errors and thus enable the user to
counterbalance them. This way, a quasi-experimental variation of the severity of errors is being made.

\[ H9 \text{ The effect of a faulty system has less impact on trust if additionally an explanation was provided.} \]

Similarly, an interaction on confidence is expected. When participants understand system errors, so they can detect and counterbalance them, they can still be confident in their decisions.

\[ H10 \text{ Participants are less confident in their decisions when they faced a faulty system, but only when no explanation was provided.} \]

As the quasi-experimental variation appears labile, a measurement for the error understanding of the participants is included. The interaction effect depends upon participants in the white box group to be more likely to understand system errors.

\[ H11 \text{ Participants facing a system with explanations report more error understanding.} \]

The second objective of this study is to investigate situational trust, which can be described as trusting single recommendations. Based on the hybrid lens model (Seong & Bisantz, 2002) it was derived that an explanation helps to match first level and second level cues. This way, wrong recommendations should be more likely to be detected, and hence less likely to be relied on.

\[ H12 \text{ Participants make less decisions in accordance to wrong recommendations, if an explanation is provided.} \]

Generally, explanations should help to assess the credibility of recommendations more accurately. This way, they are expected to enhance the performance of the human operator. Better performance can be interpreted as more appropriate trust (Llinas et al., 1998)

\[ H13 \text{ Participants in the with box group classify more stimuli correctly with the help of the system than those in the black box group.} \]

Moreover, the link between trust and reliance will be inspected. Generally, trust is believed to lead to more reliance (Lee & Moray, 1992).
H14 More trust is linked to more reliance.

Lastly, in the transfer task, participants classify stimuli on their own without aid by the system. Explanations are expected to clarify the categories so that participants are enabled to learn. A better performance would indicate a more comprehensible system (Schmid et al., 2017).

H15 Participants in the white box group classify more graves correctly in the transfer task than those on the black box group.
3 Method

3.1 Platform and Sample Size

In order to achieve a large sample size an Online Study was conducted. The effect of an explanation was conservatively estimated to be \(d = .5\) based on Dzindolet’s (2002) finding. However, the size of the hypothesized interaction effect was unclear. Using G-Power a necessary sample size of \(N = 129\) was calculated to reach sufficient power of \(1 – \beta = .8\) for finding at least medium sized effects (\(f = .25\)) in an ANOVA.

The online study was distributed through the servers of soscisurvey.de. The link was spread to students receiving credit points for participations. Additionally, it was distributed through personal channels like Facebook and Twitter. To increase motivation of participants four 5€ coupons were distributed to the four persons performing best in the classification task.

In total, the website counted 205 clicks, while 158 people completed the questionnaire. However, to preserve the quality of the data strict exclusion criteria were applied. Only data of participants who completed the whole questionnaire was analyzed. Additionally, quality measures provided by soscisurvey were used. Participants with a relative speed index above 1.6 were excluded from the sample (c.f. Leiner, 2013).

3.2 Cover Story

Participants were told the experiment was meant to address how they interact with an artificially intelligent system. Before the start of the experiment, it was shortly explained what machine learning means and that such systems could be used as decision support system for example in the field of archeology. Then, they were introduced to ship like graves that could still be found in Sweden today. Two epochs were distinguished: Iron age and Vikings age. While in both cultures the survivors would bury their deceased in ship like graves, as culture has changed so have specific characteristics of those stones. The exact time they originate from could be determined with so called radiocarbon dating, their task would be to judge the
category membership based on optical characteristics being assisted by the artificially intelligent agent.

### 3.3 Experimental Design

Each participant faced the same experimental procedure and stimuli in the same order. What differed was the capability of the system they faced, which was manipulated in a 2 (Transparency: explanation or no explanation) x 2 (Reliability: no errors, 31.25% errors) between-subject design. Participants were randomly assigned to the four conditions.

### 3.4 Stimulus material

#### 3.4.1 Stimuli

Stimulus material was designed after Medin and Schaffner (1978) with the Open Source Software Gimp. In a field of 160x160 pixels on grey (3f3f) background several black dots made with a brush of 0.75 hardness formed “graves” varying on four dimensions (see Figure 9):

- **Form**
  - 0: round
  - 1: narrow

- **Amount**
  - 0: few dots
  - 1: more than 16

- **End stone**
  - 0: thick (19.2px)
  - 1: small (9.6px)

- **Orientation**
  - 0: east
  - 1: north

![Figure 9: Example of a Viking’s Grave (1000)](image)

#### 3.4.2 Explanations

This results in 16 possible combinations and two clearly separable categories with the second dimension being irrelevant for category membership. Generally, when two of the three relevant dimensions are 0, it is a Viking’s grave – when at least two constitute 1, it is a grave from Iron Age (also see Figure 5).
The explanations, that were only presented to the explanation-condition, are constructed as conjunctions that define category membership logically. The relevant features for the recommendation were written cursive and underlined.

Graves from the Iron Age: “This grave originates from the Iron Age”

- 1xx1; Narrow ∧ North: “because it is both narrow and oriented to the north.”
- 1lx1; Narrow ∧ Small: “because it is both narrow and its end stone is small.”
- Xx11; Small ∧ North: “because its end stone is small, and it is oriented to the north.”

Graves from the Wiking Age: “This grave originates from the Viking’s Age”

- xx00; Big ∧ West: “because its end stone is big, and it is oriented to the east.”
- 0xx0; Round ∧ West: “because it is round, and it is oriented to the west.”
- 0x0x; Round ∧ Small “because it is round, and its end stone is small.”

The conjunction was translated to german everyday language with “und gleichzeitig”.

### 3.4.3 System Errors

When the recommendation of the system was false, the explanation was consistent to it (see Figure 10). The relevant dimensions causing the error was always named in the explanation. Systems in the faulty condition had an encoding error on the third dimension. They sometimes encoded small end stones as big and vice versa while the inner logical structure remained intact. Systems in the faulty condition made three errors in the Sampling and two in the following unit resulting in a total accuracy of 68,75%.

![Figure 10: A wrong Recommendation with Explanation](Image)
3.5 Experimental Task

The complete experimental procedure lasted about 15 minutes and can be divided into five general units. Every participant was presented the same stimuli. The exact order of stimuli can be found in the appendix.

3.5.1 Concept learning

After participants reported their demographic data and they received a short introduction to Machine Learning, nine items were presented in one slide in the classical 5-4 structure by Medin and Schaffer (1978) (see Figure 11). Participants were told these graves were radiocarbon dated and therefore classified correctly. They also received a hint that the graves could be distinguished solely based on optical features and additionally were told it was very important for the experiment they tried to understand what determines category membership. The forward button only appeared after 60 seconds ensuring standardized inspection time.

![Figure 11: Classical 5-4 structure from Medin & Schaffner (1978) presented to the participants for concept learning](image)

3.5.2 Sampling

The sampling comprised from two steps. First, a stimulus in 250 x 250 pixels was presented which participants were asked to classify. Second, with their click the scale disappeared and the “decision of the system” appeared below the stimulus (see Figure 12). This decision – as in the whole experiment – either contained an
explanation or not and was either right or wrong in three cases. Participants triggered the next item by clicking forward.

3.5.3 Error Feedback

After the sampling the error feedback was introduced that radio carbon dates for the graves were accessible. Those could help to determine whether the system was right or wrong. The error feedback was given on one slide with the stimuli featured on the left, the decision outlined as E or W, then – if in the white group – the given explanation and lastly on the right a check or a cross symbolizing right or wrong (see Figure 13). For standardization, the forward-button only appeared after 40 seconds.

![Decision of the System]

Figure 12: Example of a stimulus (0110) from the Sampling in the Condition without explanation.

<table>
<thead>
<tr>
<th>Grab</th>
<th>Empfehlung</th>
<th>Erklärung</th>
<th>richtig/falsch</th>
<th>Grab</th>
<th>Empfehlung</th>
<th>richtig/falsch</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>schmal &amp; Norden</td>
<td></td>
<td>✔️</td>
<td>E</td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>E</td>
<td>schmal &amp; dünner Endstein</td>
<td></td>
<td>✗</td>
<td>W</td>
<td>rund &amp; Osten</td>
<td>✔️</td>
</tr>
<tr>
<td>E</td>
<td>schmal &amp; Norden</td>
<td></td>
<td>✔️</td>
<td>W</td>
<td>rund &amp; Osten</td>
<td>✔️</td>
</tr>
<tr>
<td>W</td>
<td>rund &amp; Osten</td>
<td></td>
<td>✔️</td>
<td>W</td>
<td>dicker Endstein &amp; Osten</td>
<td>✔️</td>
</tr>
<tr>
<td>E</td>
<td>dünner Endstein &amp; Norden</td>
<td></td>
<td>✗</td>
<td>W</td>
<td>dicker Endstein &amp; Osten</td>
<td>✔️</td>
</tr>
<tr>
<td>W</td>
<td>rund &amp; Osten</td>
<td></td>
<td>✔️</td>
<td>E</td>
<td></td>
<td>✔️</td>
</tr>
</tbody>
</table>

Figure 13: Error Feedback of the condition explanation and error (left) and no explanation and no error (right)
3.5.4 Decisions with Recommendation

After inspecting the error feedback and filling out the manipulation checks, participants were instructed that it was now their turn to classify seven graves with the help of the system. This block, again, can be described as separated into two stages. First, they saw the stimulus on the left-hand side and simultaneously the system’s recommendation on the right-hand side. The two options for their answer were presented beneath (see Figure 14). With the submission of their answer, the next slide appeared where they were asked to state their confidence in their decision on a 7-point Likert scale (see 3.6.2.).

![Figure 14: Screenshot from the unit "Decisions with Recommendation"](image)

3.5.5 Transfer

After filling out the trust questionnaire, the last task was to classify graves without the help of the system. Participants were presented seven items sequentially and conducted their answer similarly. This ought to indicate whether the participants learned about the concepts during the experiment. After each decision they stated their confidence in their decision.
3.6 Self-Reports

All questionnaires which were adopted from other researchers were translated to German and backwards to English to ensure quality. The scale for pleasure and Error Understanding were self-designed. All items can be found in the Appendix.

3.6.1 Trust

Trust was operationalized by Jian et al.’s (2000) trust scale. Participants filled out the questionnaire on a seven-point Likert scale ranging from “not at all” to “absolutely” after the unit Decisions with Recommendation. The twelve items were presented in randomized order.

3.6.2 Confidence

Confidence was operationalized as a one-dimensional construct measured with a 7-point Likert scale where the extrema were linguistically anchored with “just guessing” to “absolutely sure” like other researchers did (Gigerenzer, Hoffrage, & Kleinbölting, 1991). The scale appeared after each classification on a new slide (see Figure 15).

![Figure 15: Scale for Confidence](image)

3.6.3 Dimensions of Trust

The dimensions of trust, process, performance and purpose, as theorized by Lee and Moray (1992), were operationalized through selected items from subscales from Madsen and Gregor (2000). Perceived reliability was supposed to measure performance, perceived understandability targeted process, and purpose ought to be operationalized through faith. The argument that these constructs correspond, is relatively straight forward (c.f. Lee & See, 2004). The three highest-loading items were chosen respectively and translated to German. Participants indicated their agreement to the statements on a 7-point Likert scale ranging from “not at all” to “absolutely”. The items were presented in randomized order after the error feedback.
3.6.4 Enjoyment
The scale for enjoyment during the task was specifically designed for this purpose. Items were based on words addressing positive emotions used in the BSKE (Janke, Hüppe, & Erdmann, 1999) like joy or positive mood. Participants indicated their agreement with the three statements on a 7-point-Likert scale after filling out the trust questionnaire.

3.6.5 Error Understanding
After participants reflected upon the Error Feedback, they filled out the questionnaire of Error Understanding, only if they were in the error condition. The items generally addressed the understanding of the errors covering a dimension of recognizing false recommendation and one of competence, i.e. knowing the correct answer independently. The three items were rated on a 4-point scale. Participants, who have stated a general agreement to all three items – i.e. at least 3 points on every item – will be treated as if they understand the errors.

3.7 Statistical Analysis
Personal data of participants was stored separately from the experimental data. Data was prepared with Excel calculating means and the number of correct decisions. Data analysis was conducted with IBM SPSS and R. Level of significance was set on $p = .05$. For directed hypotheses one-tailed $p$-values are reported. Before the analysis, all variables were checked for normal distribution. If not met, non-parametric tests or bootstrapping for two-way ANOVAS was conducted.
4 Results

4.1 Sample

After applying all criteria, data of 132 participants remained. 75 women and 57 men participated. Data of 25 employed people and 107 students, of whom 72 were enrolled in a psychology major and 35 in other degrees was analyzed. Mean age was 24.61 (sd = 8.42) ranging from 18 to 54. Employment status and gender were not significantly linked to any relevant outcome variable, neither did age correlate with any trust-related measurements.

4.2 Reliability of Scales

The 3-item manipulation checks showed intermediate reliability for perceived reliability (α = .63) and faith (α = .68) while perceived understandability (α = .79) was better. The self-constructed three item-measures of error understanding showed decent reliability (α = .77), whereas the 3-item scale measuring enjoyment during the task showed good reliability (α = .82).

The two subscales of the trust questionnaire only correlated $p = -.47$, $p < .001$. Thus, both were analyzed separately, where Distrust (α = .75) and positive Trust (α = .88) showed acceptable and good reliability.

4.3 Tests on normal distribution

All dependent variables were tested for normal distribution with Kolmogorov-Smirnov tests. Positive Trust and Distrust were normally distributed in all four conditions, $p > .05$. Confidence in combined Decisions and in own Decisions were respectively normally distributed in all conditions, $p > .05$, except for the condition where the system was flawless and provided an explanation, $p = .015$. The performance scores reflecting the number of correct answers in the second and third unit were not normally distributed in any condition, $p > .05$. Here, non-parametric tests were conducted. Because of a lack of alternatives for two-way ANOVAs, here, bootstrapping was conducted.
4.4 Trust Dimensions

First, a two-way MANOVA with explanation and errors as independent variables on all three manipulation checks, understandability, perceived reliability, and faith, was conducted.

Using Pillai’s trace, a significant effect of both explanation, $V = .067$, $F(2, 126) = 3.01$, $p = .033$, $\eta^2 = .067$, and errors, $V = .085$, $F(2, 127) = 3.91$, $p = .010$, $\eta^2 = .085$, was found.

Subsequent two-way ANOVAs showed that explanations had a significant effect on the understandability of the system, $F(1, 128) = 7.017$, $p = .009$, $\eta^2 = .052$; but not significantly on faith, $F(1, 128) = 3.235$, $p = .074$, $\eta^2 = .025$; nor perceived reliability, $F(1, 128) = 1.136$, $p = .289$, $\eta^2 = .009$. The effect of errors conducted by the system on perceived reliability was not significant, $F(1, 128) = 2.470$, $p = .118$, $\eta^2 = .019$; but on both faith, $F(1, 128) = 4.589$, $p = .034$, $\eta^2 = .035$; and perceived understandability, $F(1, 128) = 9.495$, $p = .003$, $\eta^2 = .069$. Participants reported more faith in correct systems ($M = 3.786$; $sd = 1.21$) than in faulty systems ($M = 3.343$; $sd = 1.222$), $t(130) = 2.093$, $p = .038$. Additionally, the interaction between errors and explanation on perceived understandability was significant, $F(1, 128) = 4.796$, $p = .030$, $\eta^2 = .036$ (see Figure 16). The interactions on faith, $F(1, 128) = 0.665$, $p = .416$, $\eta^2 = .005$, and perceived reliability were not significant, $F(1, 128) = 1.220$, $p = .271$, $\eta^2 = .009$.

![Figure 16: Semi-disordinal interaction on perceived Understandability](image)
4.5 Trust

As the subscales of Jian et al.’s questionnaire (2002) correlated only moderately both were analyzed separately with a MANOVA examining the effect of the two variations. Levene test was insignificant, $p > .05$, as was box-test, $p = .571$. Using Pillai’s trace, there was a significant effect of explanation on the dependent variables, $V = .056$, $F(2,127) = 3.76$, $p = .026$, $\eta^2 = .056$. The effect of reliability was also significant, $V = .125$, $F(2,127) = 9.07$, $p < .001$, $\eta^2 = .125$. The interaction was not significant, $V = .13$, $F(2,127) = .847$, $p = .431$, $\eta^2 = .013$.

The MANOVA was followed up with discriminant analysis, which revealed two discriminant functions. The first explained 74.0% of the variance, canonical $R^2 = .375$, whereas the second explained only 26.0%, canonical $R^2 = .233$. In combination these discriminant functions significantly differentiated the treatment groups, $L = 0.812$, $\chi^2(6) = 26.59$, $p < .001$. Removing the first function indicated that the second function alone did significantly differentiate the groups, $L = 0.946$, $\chi^2(2) = 7.16$, $p = .028$. The correlations between outcomes and the discriminant functions revealed that the first function mainly correlated with positive Trust ($r = .91$) and only slightly with Distrust ($r = -.10$). The second function loaded mostly on Distrust ($r = .995$), but also on positive Trust ($r = -.42$).

Figure 17: The discriminant function plot
As you can see in Figure 17, the discriminant function plot shows that the first function, which the Chi²-Test suggests is sufficient for significantly discriminating the groups. It mainly differentiates the 1-group from the other three. The second function additionally differentiates the 3-group from the other groups (also see Figure 18 and 19).

![Figure 18: Positive Trust in different groups; error bars indicate the standard error of the mean (SEM)](image)

More reported pleasure during the task was linked to more positive trust, $r = .22$, $p = .007$, but not significantly to Distrust, $r = .001$, $p = .992$. A multiple regression with “forced entry” was conducted to test whether the three dimensions of trust
could predict positive Trust. Requirements were checked and met. Faith, perceived reliability, and understandability together explained 25.2% of the variance, \( R = .502, F(3, 128) = 14.402, p < .001 \). Faith (\( \beta = .29, p = .001 \)) and perceived reliability (\( \beta = .25, p = .011 \)) significantly predicted the score of positive trust, but not perceived understandability (\( \beta = .08, p = .402 \)).

Adding reported enjoyment to the other two predictors added further 2.5% of explained variance (\( R = .526, F(3, 128) = 16.362, p < .001 \)). Also the corrected \( R^2 \) increased from \( R^2 = .235 \) to \( R^2 = .260 \). Faith (\( \beta = .29, p = .001 \)), perceived reliability (\( \beta = .24, p = .012 \)) and reported enjoyment (\( \beta = .17, p = .030 \)) significantly predicted the score of positive trust, but not perceived understandability (\( \beta = .06, p = .518 \)) (standardized coefficients are reported). However, perceived understandability also correlated with positive trust, \( r = .31, p < .001 \). Moreover, positive Trust correlated with the number of correct decisions with recommendation with \( r = .22, p = .011 \).

A linear regression with “forced entry” with the independent variables positive trust and confidence in own decisions on the confidence in decisions with recommendation was conducted. Durbin-Watson was 1.804; VIF = 1.055. Both positive Trust (\( \beta = .10, p = .047 \)) and confidence in own decisions (\( \beta = .80, p < .001 \)) significantly predicted confidence in decisions with recommendations with \( R = .826 \), canonical \( R^2 = .682 \).

### 4.6 Decisions with Recommendation

A t-test comparing the white box group (\( M = 5.60, sd = 1.171 \)) to the black box group (\( M = 5.22, sd = 1.270 \)) revealed a significant effect in the number of correct decisions with recommendation, \( t(130) = 1.81, p = .037 \) (also see Figure 20).

A two-way ANOVA on the mean confidence in the decisions reached with the aid of the system revealed a significant effect for reliability, \( F(1, 128) = 4.374, p = .038, \eta^2 = .033 \), but not for explanation, \( F(1,128) = 0.035, p = .851, \eta^2 < .001 \), and the interaction, \( F(1,128) = 0.315, p = .576, \eta^2 = .002 \). A t-test revealed the direction of the error effect; participants facing a faulty system (\( M = 4.756, sd = 1.054 \)) were less confident in their decisions than those with correct systems (\( M = 5.128; sd = 0.948 \)), \( t(130) = 2.13, p = .018 \) (see Figure 21). The mean confidence,
participants reported correlated with the number of correct decisions they made, \( r = .21, p = .018. \)

![Figure 20: Number of correct decisions with recommendation; error bars indicate SEM](image)

![Figure 21: Confidence in decisions with recommendation, error bars indicate SEM](image)

In the error group only, participants facing a system with and without an explanation were compared how likely it was that they relied on the two wrong recommendations in this unit. Participants relied on wrong recommendations less, when an explanation was included (\( M = 0.839, sd = 0.638 \)) than without explanation (\( M = 1.088, sd = 0.570 \)) without reaching standard thresholds of significance, \( t(63) = 1.67, p = .051. \)
4.7 Transfer Task

Participants in the white box group ($M = 5.38; sd = 1.22$) performed better than the black box group ($M = 5.13; sd = 1.25$) (see Figure 22), however, a t-test revealed the effect of $d = .20$ to be non-significant, $p = .124$.

Regarding the reported confidence in the decisions in the transfer task, a two-way ANOVA revealed no significant effects. Neither the effect of explanation, $F(1, 128) = 0.093, p = .761, \eta^2 = .001$, nor for errors, $F(1,128) = 3.234, p = .074, \eta^2 < .025$, nor the interaction, $F(1,128) = 0.011, p = .918, \eta^2 < .001$, were significant.

![Figure 22: Number of correct decisions in the transfer task; error bars indicate SEM](image)

4.8 Explorative Data Analysis

The mean Error understanding correlated with the number of correct decisions they made with recommendations, $r = .32, p = .009$, with the confidence they reported to these decisions, $r = .365, p = .003$, and with the number of correct decisions in the transfer task, $r = .353, p = .004$.

Two groups were separated with the cut-off of 9 summated points as described in the methods section. Overall, 51 participants were assigned to the group, which was treated as if they did not understand the errors, and 14 were treated as if they did. A t-test revealed, that participants with error understanding ($M = 5.541; sd = 0.979$) reported significantly more confidence in decisions with recommendations than ones without error understanding ($M = 4.541; sd = 0.977$), $t(63) = 3.393, p = .001$. 
Comparing participants with error understanding with participants, whose system did not err ($M = 5.128\ sd = 0.9477$), revealed no significant difference in how much confidence they reported, $t(79) = 1.475$, $p = 0.144$. Comparing the no-error group to those without error understanding revealed a significant difference in how much confidence they reported in decisions with recommendations, $t(116) = 3.288$, $p = 0.001$ (see Figure 23).

The mean Error understanding also correlated with the enjoyment, $r = .40$, $p = .001$, and the Distrust participants reported, $r = .253$, $p = .042$.

![Figure 23: Comparison of different groups in their confidence in combined decisions; error bars indicate SEM; * indicates significant differences in t-tests](image)

Regarding trust, participants with error understanding ($M = 3.540;\ sd = 0.936$) reported more positive Trust than their counterparts ($M = 3.363;\ sd = 1.013$), but not significantly, $t(63) = 0.729$, $p = .469$. Participants facing no system errors reported clearly more trust ($M = 4.186;\ sd = 1.015$).
5 Discussion

5.1 General Findings

The present experiment consists of a comprehensive examination of the human-computer interaction with an artificially intelligent agent in a classification task. Participants interacted with an AI decision support system, which was varied in a 2 (Transparency: explanation, no explanation) x 2 (Reliability: 31.25% errors, no errors) between-subject design. The error manipulation was conducted through an error feedback given to the participants after the sampling. Several trust-related constructs were measured through self-report and the decisions participants made in the classification task were registered as performance measures. Data from a wide-ranging sample of 132 participants was analyzed.

The largest effect in the present study is the effect of errors. Errors by the system minored trust and confidence respectively. However, it was moderated by the reported error understanding. Participants who did not comprehend the system’s errors were less confident in their decisions than those understanding them, and those whose system did not err. Manipulations effects on the trust scale were examined with a MANOVA followed up by a discriminant analysis, which is discussed in length. Errors minored trust, whereas explanations increased trust. Besides, the study finds that explanations help to enhance the efficiency of the human-computer interaction. Participants whose system provided explanations classified more graves correctly, which indicates more appropriate trust. Thus, the trust model proposed in the theory section is adopted. Contrary to expectation, explanations do not seem to enhance trust in general, but instead improve the appropriateness of trust.

5.2 Separating the Trust Scale

Contrary to H3 and the considerations of the authors of the trust scale (Jian et al., 2000), the questionnaire’s subscales were only correlated with $r = -.49$. This does not allow to conclude that both constructs constitute the ends of one underlying continuum, as proposed by the authors. In fact, an explorative factor analysis suggests a two-factor structure (see Appendix B-3). This finding is in accordance with those of other researchers who found a two-factor structure (Seong & Bisantz, 2008;
Seemingly, positive Trust and Distrust are distinct, but connected constructs. Due to this, both were analyzed independently. When positive Trust is discussed, it corresponds to the construct of trust defined in section 2.3.2, but specifically refers to the operationalization through the subscale. The different denotations are chosen for clarification. Separating both obliges to define Distrust. Based on the items, Distrust might be specified as a suspicion of the actions of an opposite and thus considering its recommendations carefully. Interestingly, having defined (positive) Trust as an attitude towards the helpfulness of another, it becomes clear that positive Trust and Distrust are not mutually exclusive. Both might be high in some individuals as one might asses others as helpful but only under the condition if one checks their recommendations cautiously (Llinas et al., 1978). As you can see in Figure 24, different qualitative types of attitude can be distinguished (cf. Benamati, Serva, & Fuller, 2006).

Confirmatory validation for the subscale of positive Trust could be found through a strong link to the subscales from Gregor and Madsen’s trust questionnaire (2000) (also see 4.4). Additionally, it was closely linked to confidence in decisions with recommendation. However, its main scope of validation was unsuccessful as it could not predict reliance (see 5.8).

### 5.3 Positive Trust and Distrust

Two-way MANOVAs with explanation and error as independent variables on positive Trust and Distrust as dependent variables, and, secondly, on the three
dimensions of trust were conducted. MANOVAs are recommended for analyzing a set of variables which are theoretically connected, and for preventing Type-1 error inflation (Bray & Maxwell, 1982). However, effects found in MANOVAs are hard to interpret and there is no consensus on how to conduct the successive step of explaining group differences after identifying significant effects (Bray & Maxwell, 1982). For effects on the trust dimensions, subsequent ANOVAs were chosen since the specific effects of the manipulation are of interest (see 5.5). However, positive Trust and Distrust were previously postulated as one construct, and thus, were analyzed with a discrimination analysis, as suggested for this purpose (Borgen & Seling, 1978).

The discrimination analysis identified two factors. Canonical variate correlations are reported for interpreting the factors as they have been found to be more stable in cross-validation than discriminant coefficients (Dunca & Timm, 1978). The first function clearly loads onto positive Trust, and the second strongly on Distrust, but also moderately on positive Trust. Inspecting the discriminant function plot, it becomes apparent that function 1 alone successfully discriminates the data, as indicated by a significant Wilks-Lambda. Note that function 1 is easily interpretable, as it almost only loads on positive Trust. It sorts the four groups in the order of 1 (explanation and no error), 3 (no explanation and no error), 2 (explanation and error) to 4 (no explanation and error) (see Figure 25).

![Figure 25: Schematic representation of how the first canonical function discriminates between the experimental groups indicating both significant effects; for the exact discriminant function plot see Figure 15](image)

This way, it mainly maps the error effect through the visible gap in between the two pairs. However, the effect of explanations is visible, too, as the group 2 lies ahead of 4, and 1 ahead of 3. These effects are not significant in individual t-tests, however, collectively a significant effect for explanation was found in the MANOVA. As noted, factor 1 is sufficient for discriminating the groups. Regarding
factor 2, nonetheless, it can be seen that it mainly discriminates group 3 from the others. Inspecting the means suggests that this phenomenon originates in a visible drop in Distrust in this group. Apparently, participants whose system did not err and was not faulty, established less Distrust than others. This is graspable in so far that suspicion in perfect systems is needless, but in contrast to group 1, it is also harder to develop, when no explanation is offered to scrutinize recommendations.

Conclusively, it can be reiterated that both manipulations had a significant effect on the measures of positive Trust and Distrust. Confirming H1 and H2, errors minored positive Trust, whereas explanations increased positive Trust, however, with slighter effect than errors. Distrust, on the other hand, did not differ much between the groups except for participants facing a correct system with explanations.

Positive Trust was linked to the reported enjoyment during the task with a correlation of $r = .22$ supporting H8. Likewise, satisfaction has been shown to be positively correlated with trust in both vehicle automation (Donmez et al., 2006) and combat identification (Wang et al., 2011). Yet, correlative findings cannot be interpreted as causal relations. It is equally plausible that trusting more generates more enjoyment or a third variable influences both just like the postulated idea that perceived positive emotions help to build trust. However, a recent study by Stokes et al. (2010) provides such causal evidence. They induced a positive or negative mood in participants of two groups, respectively, through the International Affect Picture System (Lang, Bradley, & Cuthbert, 2008). In fact, the study finds a significant difference between the groups in how much trust they reported after an initial trust formation. Lee and See (2004) first proposed that trust would be mainly shaped emotionally. They suggest the user interface to be the main basis for communication and thus to be especially important for shaping a successful interaction and thereby trust. This suggests that focusing on an enjoyable usability helps to enhance trust. But also task-related aspects, like the ease of task fulfillment might be an mediator to both enjoyment and trust. When a system simplifies a task fulfillment, the task is perceived as more pleasant, and the system would be trusted more.

However, H14, postulating that more trust is linked to more reliance, could not be confirmed. The number of occasions when the participant chose the option the system recommended were counted and analyzed. Positive Trust could not predict this number. This is surprising in so far, that the main objective for the construct of trust,
is predicting reliance (Llinas et al., 1998). However, multiple researchers have concluded similar inability to find a coherent link (Wiegmann, Rich, & Zhang, 2001; Lee & Moray, 1992). This finding aligns with the idea of separating learned trust and situational trust. The former rather refers to a global assessment of the opposite, whereas the second one accords to trusting single recommendation. This construct would accord to confidence in a recommendation, which was directed to assess through the hybrid lens model. This would assumingly quite ultimately result in reliance. However, this construct was not measured on the present research, which leaves this question open for future research. Instead, positive Trust and the performance in the classification task with the aid of the system correlated significantly with $r = .22$. Again, as a correlative finding this can be interpreted in both ways. Possibly, higher trust leads to better performance which would demand to build as much trust as possible. On the other hand – and theoretically more meaningful – could those who perform better, attribute this circumstance on the AI’s helpfulness, which then again led to more trust.

### 5.4 Interaction on Trust and Confidence

Based on several theoretical considerations, an interaction between errors and explanation on trust was proposed. However, such interaction was found neither in the MANOVA, nor in an ANOVA on positive Trust contradicting H9. Likewise, H10 proposing an interaction on the confidence participants report in decisions with recommendations was disconfirmed. This lack of evidence is mainly attributable to the fact that participants did not understand the errors, which would have been a necessary condition for the effect. Contrary to H11 participants in the white box group did not report more error understanding. Moreover, the interaction on perceived understandability indicates that participants were confused by system errors. Both are, in part, due to a smaller effect of explanation than previously expected (see 4.6).

The quasi-experimental manipulation expecting the white box group to grasp the structure of system errors, which was supposed to lead to an interaction, was unstable and hence problematic from the beginning. This is why the measurement for error understanding was included in the questionnaire. On the one hand it backs up the claim why the effect was not found, on the other hand it allows a post-hoc
analysis with error understanding as a quasi-experimental variation. As error understanding was postulated as a dichotomous construct, participants were divided into two groups with a threshold of nine summated points. Participants scoring at least nine points summed-up on the scale with a 4-point Likert scale, indicated general agreement to all three items. The measurement of error understanding demonstrates validity as it positively correlated with the number of correct decisions in the classification task with recommendations.

Analyzing the means of confidence in decisions with recommendations reveals a moderating effect of error understanding. Those assigned to the error understanding group reported significantly more confidence in their decisions than those not understanding the errors – but also more than those not facing system errors. The first finding makes sense in so far, that knowing about defects in the system enables the user to counterbalance them. Being aware of such understanding, allows to be confident in the decisions regardless possible errors of the system. The second finding indicates that system errors, when understood, even increased confidence slightly. Due to the small number of participants in the error understanding group, this finding is insignificant and should be interpreted cautiously.

The general framework provided by Siegrist et al. (2003) separating trust and confidence was supported with a linear regression. The basic idea was that confidence arising after a decision with recommendation was made either stems from confidence in own competences, which were measured with confidence in decisions in the transfer task, or from trust in the system, operationalized through positive Trust. In fact, both significantly predicted the confidence participants reported with an $R^2 = .682$. This is an unusually high percentage of explained variance strongly supporting the theoretical considerations.

### 5.5 Trust Formation

The dimensions of trust were measured through 3-items adopted from subscales from Madsen and Gregor (2000) after the sampling and the error feedback.

The interaction between errors and explanation on the perceived understandability is one of the most interesting findings of this study (see Figure 16). On the one hand because it is contrary to H4 expecting an effect of explanation only, on the other because it is intuitive and gives insight into why other hypothesis might not have
been confirmed. A semi-disordinal interaction was observed: participants reported the system only to be more understandable when an explanation was provided, and when it made no errors at the same time. This indicates that the systems did not appear comprehensible to participants when it simultaneously was faulty. Apparently, participants were not able to understand these errors but were rather confused. Transparency, however, was the hypothesized mediator on trust. This might explain the unexpectedly small effect of explanation on trust (see 5.6), and, as well, why no interaction between errors and explanation was observed (see 5.4).

Contrary to H5 errors had no effect on participants’ perceived reliability. However, this might be due to its items, which showed the worst internal consistency with $\alpha = .63$ in the present experiment, just like in Madsen and Gregor’s (2000) original paper. However, the error variation had a significant effect on the faith participants reported in the systems, contradicting H6. This finding is contrary to expectation, because this dimension was not systematically varied. Apparently, participants perceived the purpose of the system as less benevolent when it erred. This can be explained by derivations between the different dimensions (see 2.3.3).

Moreover, confirming H7, the three measurements of the trust dimensions were closely linked to positive Trust. All measurements of the dimensions positively correlated with positive Trust (see Appendix B-2). Besides, they explained 25.2% of its variance as shown in a linear regression with “forced entry”. However, perceived understandability did not contribute to the prediction significantly. Adding reported enjoyment to the model revealed its significant contribution to the prediction, which increased the explained variance by 2.5%. This indicates that the emotional facet contributes to another level of description which might be added to the theory.

Overall, the general validity of Lee and Moray’s (1992) idea of the three dimensions of trust was reinforced. However, comparing them against each other, indicates that the understandability of the system plays a smaller role in the trust formation than purpose and performance of the system. Yet, it is unclear whether this is due to the theoretical construct of understandability or its operationalization. Whereas some have suggested that a system’s understandability itself might shape trust (Lee & See, 2004), this does not coincide totally with Lee and Moray’s descriptions. They suggested that knowing about the inner workings and being confident they adequately function leads to trust. Yet, this finding is surprising, as transparency was
expected to be a main path to the trust formation. This is further discussed in section 5.6, resulting in an adoption of the trust model in section 5.7.

### 5.6 Small Effect of Explanation

Before the experiment, the effect of explanation on trust was estimated with $d = .5$ ($\eta^2 = .06$) based on a similar finding by Dzindolet et al. (2003). One the one side, the effect size from the MANOVA almost perfectly corresponds to this, $\eta^2 = .056$. On the other side, in the ANOVA, the effect of explanations on positive trust was only $\eta^2 = .012$ in the expected direction (see Appendix B-1). Other hypotheses, dependent on the explanation-variation were contradicted. Contrary to H15, participants in the white box group did not perform significantly better in the transfer task than their counterparts, although the tendency revealed the expected direction. This suggests that the system with explanations was not (significantly) more comprehensible (Schmid et al., 2017), which, then again, is supported by the finding of an interaction on perceived understandability (see 5.5.).

One explanation for why the effect was smaller than anticipated, may be the error feedback. All stimuli, recommendations and their correctness were presented in one slide. This way, they received information how well the system performed similar to the meta-information used by Seong and Bisantz (2008). They found that such meta-information affects the dimension of process as it enhances the transparency of the system. It minored trust for highly reliable systems but increased trust in faulty systems, somewhat converging both. This was interpreted in such a way that the feedback helped to remind participants that good systems were not perfect, but faulty systems were neither catastrophic resulting in more adequate expectations (Hoff & Bashir, 2015). As this error feedback slide was present in all four conditions, it is impossible to segregate its effect here. It seems plausible that the error feedback has increased the transparency of the system itself which means no participant faced a “truly” black box. This might have rendered further opportunities for transparency unnecessary, especially when the system performed flawless.
5.7 Adopted Trust model

Conclusively, an adoption of the presented trust model should be considered. Based on this study’s findings and Seong and Bisantz’ (2008), one might argue that transparency is no source of trust itself but rather a potential path to build trust. Transparency – or explanations – help to regard the true capabilities of the opposite and thereby assess its trustworthiness more precisely. This, however, is no ultimate path to more trust, but can also diminish trust. An example from interpersonal relationships might shed light on this. Think of a distant acquaintance you have known for a couple of months. In one of the first intimate conversations you have, he tells you that he regularly lies in his job to gain advantages and describes one occasion where he stole money from his employer. This constitutes an instance of transparency – but quite certainly will not increase the trust you put in him. Also with automation, explanations did not enhance trust but diminish it in Ribeiro et al.’s experiment (2016), because they revealed the defective basis recommendations were based upon. Transparency offers an opportunity to establish trust but also to correct for overtrust. Here the two other dimensions come into consideration: transparency helps to identify the purpose and the performance of the opposite. When they are considered positively, trust will increase. On the other hand, when defects are detected, it can also decrease. Thus, transparency is a mediator (see Figure 26).

Similarly, Ososky, Schuster, Phillips, and Jentsch (2013) differentiate between two types of trust. The first one is trust in the intention (purpose) of others, knowing they are not deceptive. Secondly, they suggest trust in abilities or competence aligning with the concept of performance. It appears probable that humans operate under

![Figure 26: Adopted Trust model with Transparency as a mediator to Trust](image-url)
this general scheme somewhat aligning with a theory of mind (Meltzoff, 1995): intentions cause actions, which then again have consequences for myself.

5.8 Appropriate Trust and Reliance

This idea, that transparency – or explanations – do not necessarily build trust, but rather help to build appropriate trust, is supported by another main finding of the present experiment.

Consistent with H13, participants in the white box group made more correct decisions with recommendation than their counterparts. This finding aligns with others indicating that the usage of information automation is more robust to system errors than decisions automation (see 2.5.1). As a theoretical foundation, the hybrid lens model was applied to analyze the decision of an operator in a classification task being provided with a recommendation (Seong & Bisantz, 2008). Here, different levels of cues can be differentiated. First, the ones from the stimulus, and secondly, the ones from the automation which provides the recommendation can be distinguished (see 2.5.2). The importance of explanations is illustrated as they connect second to first level cues. An explanation embeds a recommendation into the context on which it is based. This way, operators are enabled to judge its credibility. They can match the content of an explanation with the cues they observe themselves. If they match, the recommendation is likely to be trustworthy. If they do not, the recommendation appears to be inadequate to the context the users perceive, and they are less likely to rely on it. That wrong recommendations with explanations are less likely to be relied upon was also derived to H12. In fact, participants whose systems provided an explanation, were 12.4% less likely to take the same decision as the system recommended, when it was false. This finding with \( p = .051 \) does not meet the level of significance and should be interpreted carefully. Due to its tightness, however, it seems worth reporting.

This finding that explanations help to enhance the performance of the human-automation team is highly relevant for practical implementations. For practitioners of Machine Learning in all fields of application, obviously, the final outcome is of interest. When constructing automation, they should consider the human operator that oversees the automation. Slightly lower accuracy in the automation might be bearable, when, instead, the system is interpretable to humans.
This finding should be of major interest for theoretical research as well. As noted by Llinas et al. (1998), better performance can be interpreted as more appropriate trust. Seemingly, a user, who performs better with the automation, has judged its helpfulness more appropriately. Thus, it can be derived that explanations help to shape appropriate trust. Appropriate trust is one of the major subjects of research in HCI-research (Ososky et al., 2013; Lee & See, 2004).

5.9 Limitations

One major limitation which should be considered when interpreting the results is that it was implemented as an online experiment, where confounders are harder to exclude. It is plausible that participants conducted the experiment under different conditions. Furthermore, some might have rather “clicked through” the questionnaire which is a common problem. For this purpose, however, strict exclusion criteria were applied.

Secondly, it was impossible in the present design to balance the order of the stimulus material. As reported, the effect of errors also depends on their timing, which is why the error placed on the second place in the sampling might have had a strong effect.

Thirdly, the precise composition of the experiment should not be overlooked. The error feedback, with which the manipulation was achieved, constitutes a very special form, which will be barely found in the real world. There, most errors will be encoded “live”, which consists of different cognitive mechanisms. Among others, it requires background knowledge to detect errors in the first place, which was one reason not to implement it in the experiment as it would have generated more noise.

Lastly, no evidence is irrefutable. Any finding should be interpreted with caution until further experiments replicate them. It should also be noted that one-tailed p-values are reported for directed hypotheses. Although this is not uncommon, it can inflate Type-I errors.
5.10 Conclusion

This paper is theoretically based upon a separation of learned trust and situational trust. Referring to the first, errors minored trust and explanations increased trust. Secondly, the reliance is considered under the concept of situational trust. Participants performed better with the aid of the system, when an explanation was provided, which indicates more appropriate trust.

Secondly, the provided literature review offers points of reference for future research. Lee and See’s (2004) definition of trust, but also Lee and Moray’s (1992) dimensions of trust are advised to be considered when researching trust in automation. Besides, a modification of the trust model from Lee and Moray is proposed, suggesting that transparency is no ultimate path to more trust. Instead, it offers an opportunity to assess the credibility forming more appropriate trust.

Lastly it is advised that researchers should further strive for comprehensible approaches in machine learning. Sole focus on accuracy is insufficient, and the perspective of the human operator should be considered. Approaches like LIME (Ribeiro et al., 2016) or LIME-Aleph (Rabold, Siebers, Schmid, 2018) combining interpretable approaches on top of neural nets should be further investigated to ultimately “break the black box”.

6 References


Appendix A: Experimental Material

A-1 Stimulus Material
A-2 Questionnaire: Dimensions of Trust (Madsen & Gregor, 2000)
A-3 Questionnaire: Trust Scale (Jian et al., 2000)
A-4 Questionnaire: Enjoyment and Error Understanding
A-1 Stimulus Material

Concept Learning

1111; 1011; 1011; 1101; 0111

1100; 0110; 0001; 0001

Sampling

1001; 1000; 1111; 0010; 0101; 0011; 0100; 0110, 1010

Decisions with Recommendations

1011; 0001; 0000; 1100; 1101; 0111; 1110

Transfer Task

0010; 0100; 1001; 1000; 0101; 1111; 0011

Note. Based on the two bold numbers the explanations were based. For stimuli written cursive, errors occurred in the respective group. This is why three numbers are written bold for error-items.
A-2 Questionnaires: Dimensions of Trust (Madsen & Gregor, 2000)

Perceived Understandability

_U1_ Ich weiß, wie sich das System beim nächsten Grab verhalten wird, weil ich weiß, wie es funktioniert.

_U2_ Auch wenn ich nicht genau weiß, wie das System funktioniert, weiß ich, wie ich es benutzen muss, um Entscheidungen zu treffen.

_U3_ Ich verstehe, wie das System mich unterstützt, Entscheidungen zu treffen.

Perceived Reliability

_R1_ Das System verhält sich reliabel.

_R2_ Das System analysiert Probleme auf eine konsistente Weise.

_R3_ Ich kann mich darauf verlassen, dass das System richtig funktioniert.

Faith

_F1_ Ich folge Empfehlungen des Systems, sogar wenn ich mir nicht sicher bin, ob sie richtig sind.


_F3_ Sogar wenn ich keinen Grund habe, zu erwarten, dass das System ein schwieriges Problem lösen kann, fühle ich mich sicher, dass es das wird.
A-3 Questionnaires: Trust Scale (Jian et al., 2000)

D1 Das System ist trägerisch.
D2 Das System verhält sich auf eine hinterhältige Weise.
D3 Ich bin misstrauisch, was die Absicht, das Handeln und das Ergebnis des Systems betrifft.
D5 Die Empfehlungen des Systems haben schädliche Resultate.
T1 Ich bin zuversichtlich in die Empfehlungen des Systems.
T2 Das System bietet mir Sicherheit.
T3 Das System ist integer.
T4 Das System ist zuverlässig.
T5 Das System ist reliabel.
T6 Ich kann dem System vertrauen.
T7 Ich bin mit dem System vertraut.

A-4 Questionnaires: Enjoyment and Error Understanding

Enjoyment

E1 Während der Klassifizierung der Gräber war ich gut gestimmt.
E2 Während der letzten Minuten war ich gut gelaunt.
E3 Wenn ich ein neues Grab klassifiziert habe, empfand ich Freude.

Error Understanding

E1 Ich habe verstanden, wann das System fehlerhafte Vorhersagen macht.
E2 Ich erkenne, wenn die Vorhersage des Systems falsch ist.
E3 Wenn das System falsch liegt, kenne ich trotzdem die richtige Antwort.
Appendix B: Additional Results

B-1 Separate ANOVAS on positive Trust and Distrust

B-2 Table 1: Correlations of Trust-related self-reports

B-3 Table 2: Factor Analysis of the Trust scale
B-1 Separate ANOVAS on positive Trust and Distrust

Separate univariate two-way ANOVAs on the outcome variables revealed only a significant effect of errors on positive trust, \( F(1, 128) = 17.82, p < .001, \eta^2 = .122; \) its effect on Distrust was not significant, \( F(1, 128) = 2.36, p = .127, \eta^2 = .018 \). Explanation had no significant effect on Distrust, \( F(1,128) = 2.33, p = .13, \eta^2 = .018 \); nor on positive Trust, \( F(1,128) = 1.49, p = .224, \eta^2 = .012 \). Neither the interaction on positive Trust, \( F(1, 128) = 0.069, p = .793, \eta^2 = .001 \); nor Distrust was significant, \( F(1, 128) = 1.535, p = .218, \eta^2 = .12 \).

B-2 Correlations of Trust-related self-reports

Table 1. Correlations of trust-related self-reports

<table>
<thead>
<tr>
<th>Variable</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Positive Trust</td>
<td>-.49**</td>
<td>.31**</td>
<td>-.42**</td>
<td>.42**</td>
<td>.22**</td>
</tr>
<tr>
<td>2 Distrust</td>
<td>.09</td>
<td>-.04</td>
<td>-.07</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>3 P. Understandability</td>
<td>.31**</td>
<td>.56*</td>
<td></td>
<td>.14</td>
<td></td>
</tr>
<tr>
<td>4 Faith</td>
<td></td>
<td>.44**</td>
<td></td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>5 P. Reliability</td>
<td></td>
<td></td>
<td></td>
<td>.11</td>
<td></td>
</tr>
<tr>
<td>6 Enjoyment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Explanatory Note.* P. = Perceived. * \( p < .05 \), ** \( p < .01 \).
### B-3  Factor Analysis of the Trust scale

Table 2. Factor loadings of items of the trust scale (Jian et al., 2000) in an exploratory factor analysis with Varimax

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 Deceptive</td>
<td>−.236</td>
<td>.743</td>
</tr>
<tr>
<td>02 Underhanded</td>
<td>−.207</td>
<td>.784</td>
</tr>
<tr>
<td>03 Suspicious</td>
<td>−.072</td>
<td>.668</td>
</tr>
<tr>
<td>04 Wary</td>
<td>−.406</td>
<td>.365</td>
</tr>
<tr>
<td>05 Harmful</td>
<td>−.139</td>
<td>.795</td>
</tr>
<tr>
<td>06 Confident</td>
<td>.808</td>
<td>−.080</td>
</tr>
<tr>
<td>07 Security</td>
<td>.813</td>
<td>−.105</td>
</tr>
<tr>
<td>08 Integrity</td>
<td>.581</td>
<td>−.337</td>
</tr>
<tr>
<td>09 Dependable</td>
<td>.788</td>
<td>−.280</td>
</tr>
<tr>
<td>10 Reliable</td>
<td>.730</td>
<td>−.063</td>
</tr>
<tr>
<td>11 Trust</td>
<td>.696</td>
<td>−.442</td>
</tr>
<tr>
<td>12 Familiar</td>
<td>.572</td>
<td>−.153</td>
</tr>
</tbody>
</table>

*Explanatory Note.* Items are named after their main adjective. Complete items and their translation can be found in Appendix A.
Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die vorgelegte Arbeit selbständig angefertigt, dabei keine anderen Hilfsmittel als die im Quellen- und Literaturverzeichnis genannten benutzt, alle aus Quellen und Literatur, einschließlich des Internets, wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht und auch die Fundstellen einzeln nachgewiesen habe.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Bamberg, 21.08.2018

Johannes Markus Burr