University of Zurich[UZH]

Thesis August 22, 2019

When the Turing Test Meets Trust

Comparing Human and AI Explanations

Florian Patrick Ruosch
of Kloten ZH, Switzerland

Student-ID: 11-932-290
florian.ruosch@uzh.ch

Advisor: Suzanne Tolmeijer

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
http://www.ifi.uzh.ch/ddis

Acknowledgements

First and foremost, my thanks go out to Suzanne Tolmeijer for her help, support, and guidance during this project.

Next, I would like to thank Professor Dr. Abraham Bernstein for allowing us to work on this topic as well as his supervision and valuable inputs.

Also, my gratitude to Dr. Daniele Dell'Aglio and Cristina Sarasua for their hours of brainstorming and problem solving.

Last but not least, I say thank you to my father Hans-Peter Ruosch for his support and understanding over the past few months.

Thank you all for allowing a spark of an idea to become this fully-fledged work which you are now holding in your hands.

Abstract

With the rise of AI, smart technology is taking over many aspects of our lives. We rely on it increasingly more often for simple and also for complex tasks. But do people really trust these smart systems or do they still prefer the old-fashioned human? To answer this question, this work explores trust in AI. We used a neural network as a representative and image classification as an example task that can be performed by a smart system. Is a user's trust in an answer influenced by knowing whether it was given by another human or by an AI? To check for a possible bias, we conducted an experiment in the form of a survey with 900 participants on the crowd-sourcing platform Amazon Mechanical Turk. It pitted labels for images and their visually represented explanations obtained from the neural network against those produced by humans. Using a multi-dimensional scale to measure trust, we gained insights for different settings. They varied regarding the available information: giving the origin of label and explanation versus withholding or disguising sources, e.g. a human-generated label and explanation is presented as coming from AI. We compared the results and found few statistically significant differences between the various setups. This led us to conclude that no clear bias exists toward AI- or human-produced results and that knowledge about the source and the availability thereof does not exhibit a distinct influence on trust of humans in AI.

Zusammenfassung

Mit dem Aufstieg von AI hält smarte Technologie Einzug in viele Aspekte unseres Lebens. Wir verlassen uns immer öfter darauf für einfache und auch für komplexe Aufgaben. Aber vertrauen Leute wirklich diesen intelligenten Systemen oder ziehen sie immer noch den altmodischen Menschen vor? Um diese Frage zu beantworten, beschäftigt sich diese Arbeit mit Vertrauen in AI. Wir verwenden ein neuronales Netzwerk als ein Vertreter und Bildklassifizierung als ein Beispiel für eine Aufgabe, die von einem intelligenten System übernommen werden kann. Wird das Vertrauen eines Benutzers in ein Ergebnis durch das Wissen beeinflusst, ob es von einem Menschen oder einer AI kam? Um die mögliche Existenz von Vorurteilen zu prüfen, führten wir ein Experiment in Form einer Umfrage durch mit 900 Teilnehmern auf der Crowdsourcing Plattform Amazon Mechanical Turk. Es stellte Klassifikationen von Bildern und die zugehörige visuell repräsentierte Erklärung produziert durch das neuronale Netzwerk den von Menschen gemachten gegenüber. Mittels einer mehrdimensionalen Skala zur Messung von Vertrauen erhielten wir Einblicke für verschiedene Set-ups. Diese variierten bezüglich den verfügbaren Informationen: Die Herkunft von Label und Erklärung ist gegeben oder nicht sowie Verschleiern der Quelle, wie zum Beispiel von Menschen gemachte Label und Erklärung werden als von einer AI produziert dargestellt. Wir verglichen die Resultate und stellten wenige statistisch signifikante Unterschiede fest zwischen den unterschiedlichen Konstellationen. Das führte uns zur Schlussfolgerung, dass keine klaren Vorurteile vorhanden sind bezüglich von AI oder Menschen gemachten Antworten und dass Informationen zur Quelle sowie deren Verfügbarkeit keinen eindeutigen Einfluss aufweisen auf Vertrauen von Menschen in AI.

Table of Contents

1

Introduction

In recent years, we have seen a rapid increase in the capabilities of Artificial Intelligence (AI) [LeCun et al., 2015, Krizhevsky et al., 2012]. This has also led to a wide deployment of these smart systems (used interchangeably with AI in this work). They play an integral role in our everyday lives. Their tasks range from seemingly unimpactful functions like optimizing the battery usage for our smartphones to crucial activities such as the self-driving system in the newest cars. While these two examples both simplify our lives, they hold one important difference: the consequences of their failure. Phone manufacturers may be fine with having a 1% error rate on the prediction of the usage of the smartphone which results in inefficient battery optimizations. Cars, on the other hand, may not even be permitted on the road with a 0.1% error rate because it may lead to thousands of accidents per day resulting in at least as many users injured. With these ramifications, humans are required to place their trust in these smart systems because they (at least partly) relinquish control over something. This relies heavily on transparency and explainability [Siau and Wang, 2018], which led to the term eXplainable AI (XAI) coined in [Gunning, 2017].

Even though AI may have caught up to human performance levels and even surpassed us in certain tasks, the most efficient methods are a black box [Samek et al., 2017b]. There is no way of getting a look at their inner workings which results in a lack of transparency even though it can be crucial for certain applications (e.g. medical field). An effort has to be made in the direction of interpretability and explainability.

In this work, we examine the implications of explanations in an AI setting inspired by the Turing Test [Turing, 1950], which has received much criticism when it comes to determining intelligence in AI, but it still can be relevant in the context of trust and bias toward machines. The authors of [Hayes and Ford, 1995] name two major drawbacks of the Turing Test: there is no way of recording small advancements and an impartial assessment is difficult. The former criticism says that it only checks for a total result (i.e. pass or fail) which is important in the proposed setting as trust is either gained or not. This converts the presumed disadvantage into an actual advantage. The latter argument may prove true when evaluating a system for performance (is it fast or good enough?) because judges may have seen similar systems and learned from their examinations. But trust is invoked intrinsically and its requirements may differ from person to person [Hoff and Bashir, 2015]. This invalidates the criticism, that the Turing Test needs unbiased opinions. In [Moor, 1976], the author describes an argument that the Turing Test could

be treated as flawed since a machine may pass it with some unconventional procedures which can be cause to posit that it did not think at all as required by the test. Curiosity invoked by human nature then demands insights into the inner workings of said machine. Understanding how it operates *might* change the conclusion that it thinks. As long as it remains unproven that knowledge of the inner workings are *necessary* to judge this ability, this criticism can be treated as void. In our approach to the Turing Test, we explicitly provide evidence of the internal mechanics of a machine (namely explanations) to study the effects on users. This leads to the following research question:

(RQ) Given an image classification and a visual explanation of the classification, is the user's trust in the system influenced by knowing whether the answer was given by a human or by an AI?

Because of the current, excellent state of AI for image classification and the availability of resources, we chose this task as the centerpiece of the experiment. In this context, we designed a system using an existing state-of-the-art CNN toolbox and an off-the-shelf explanation framework. Then, we conducted a study where we ask the participants questions related to trust. Does their knowledge of the source for the prediction and the explanation affect the amount of trust they place in the system?

This work is structured as follows: in Chapter 2 we discuss the definitions and effects of trust in smart systems and present an overview of solutions to explain AI with a special focus on Convolutional Neural Networks (CNN) [LeCun et al., 1998, Krizhevsky et al., 2012]. The design of the experiment and the questions we want to answer are explained in Chapter 3. Chapter 4 treats the implementation of the system that produced the data for the proposed experiments as well as how we deployed the survey. In Chapter 5, we evaluate the results. This is followed by Chapter 6 which outlines the identified limitations and points to work left for the future. Finally, we draw the conclusions in Chapter 7.

2

Related Work

In this chapter, we discuss the related work. We start out by examining what trust is in the context of XAI and its impact. We then work our way through surveys for XAI toward multiple options for general explanation systems. A special focus is put on the evolution of explainability of CNN for image classification. Last but not least, we look at the algorithms used in the implementation.

2.1 Trust and AI

First, we need to define trust before we can discuss its implications for software. The authors of [Mayer et al., 1995] may have a background in economics, but their formulation is relevant in this scope nevertheless. They define trust as:

> "[...] the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." (p. 712)

In the context of AI the *trustee* (the later-mentioned party) is the smart system (used synonymously with AI and machine in this work) and the *trustor* (the first-mentioned party) is the user.

The authors of [Verberne et al., 2012] take this definition up and state that trust in smart systems is closely related to it providing information and having the same goals as the user. Not only do these conditions lead to trust but also to acceptance.

The paper [Siau and Wang, 2018] focuses on building trust in AI, but the authors share the view that acceptance relates to trust. Furthermore, they distinguish between two types of trust: initial trust built based on cues or dispositions and continuous trust which needs to be nurtured. Each has different factors that affect trust-building. Initial trust is formed by performance which includes representation, perception, and reviews as well as characteristics from the process like transparency, the ability to explain, and trialability. Developing continuous trust is also affected by performance, but the features are usability and reliability, collaboration and communication, sociability and bonding, security

and privacy protection, and interpretability. A second factor is related to purpose with keywords being goal sharing and concerns about job replacement.

The authors conclude that trust in AI differs from that in other technologies since the process, purpose, and performance for smart systems lack clear definitions. Nevertheless, trust is essential in the evolution and the acceptance of AI.

In [Langley et al., 2017], the authors state that trust by human users requires smart systems to explain and reason about their behavior. They introduce the term *explainable agency* and characterize its task as follows: given a set of objectives and some background knowledge, they generate plans for a predefined problem, execute them (while adapting where necessary), and are required to produce records of decisions, summary reports, as well as understandable answers to questions about their choices. Therefore, they require four abilities: explaining decisions while generating a plan, reporting executed actions, reasoning why it diverged from a plan as well as adapting to actual events, and communicating its decisions and reasons.

The authors also elaborate on criteria for evaluating explainable agents. First, the subjective ratings about the quality of answers to questions in regards to clarity and suitability. Second and more objectively, how well people can predict the behavior of an agent in future situations after they interacted with the system.

2.2 Building Trust with XAI

The authors of [Samek et al., 2017b] claim that trust between AI and humans is similar to trust relationships between people for which explanation is often a prerequisite. They bring forth arguments on why explanations in smart systems are necessary. Their first reason is verification: by default, a user should not trust a black box system. Furthermore, explanations facilitate improvements. Not only is it easier to compare systems, but we can also find weaknesses and discover biases (in model or data). Another major argument is knowledge transfer, where learning from an existing system is made possible by distilling its knowledge and therefore gaining new insights. Last but not least, AI needs to comply with current and future legislation. In the European Union, a law took effect in 2018, which gives users a right for an explanation when their life is affected by a decision made by an algorithm [Goodman and Flaxman, 2017].

In the second part of their paper, the authors elaborate on a potential evaluation of the quality of explanations. They rely on a measure based on perturbation analysis proposed in [Samek et al., 2017a]. To assess an explanation introduce noise to an input variable deemed highly important for the prediction. This should lead to a steeper decline of the prediction score than perturbation of less important inputs. An objective measure for the quality of an explanation can be obtained by iteratively perturbing input variables and keeping track of the decline of the prediction score.

2.3 Interpretability and Explainability in AI

The survey [Dosilovic et al., 2018] explicitly connects explainability to trust by pointing out the newly arising problems with the advances of AI in areas such as medicine or self-driving cars. Because of the way humans interact with smart systems (and the other way around since they also affect our lives), a trust-relationship must be built. Therefore these systems must satisfy many criteria [Israelsen and Ahmed, 2019], such as explanatory justifiability, usability, fairness, or reliability. The authors then cite definitions for trust, interpretability, comprehensibility, and explainability from literature, but they end up concluding that there are no unique definitions, terms are used interchangeably by researchers, and formalization is impossible since no definition is strict enough.

They continue by presenting two methods for interpretability and explainability: integrated and post-hoc. The former relies on transparency and in turn trades off performance since they are conflicting goals [Yaochu Jin and Sendhoff, 2008, Freitas and A., 2004]. It comes in two forms; pure (restricting itself to transparent models) and hybrid, where transparent model families are combined with black-box methods. Post-hoc interpretability is not dependent on the inner workings of the model and has no effect on performance. The authors distinguish two types of post-hoc methods: those addressing interpretability and others covering explainability. Interpretability can be achieved by having a transparent proxy model approximating the prediction of the black-box or in an indicative approach with conceptual representations such as visualizations. A common form of explaining is not just having a prediction as output but also including a list with features and their significance in the decision while other methods present explanations as visualizations, text, examples, etc. The authors end up concluding that not enough studies have been conducted on interpretability with user-based metrics and point out that more focus should be put on these less explicit criteria instead of the optimization objectives.

In the survey [Zhang and Zhu, 2018] the authors call for the need to visualize Deep Neural Networks (DNN) and especially CNNs. They argue that DNNs obtain their impressive performance by sacrificing interpretability because of their black-box nature. This makes them hard to interpret apart from the final output layer.

The authors then identify five research directions to improve visual interpretability. When it comes to unraveling the combinations of patterns found in CNN representations, two interpretable solutions are described: explanatory graphs and decision trees. Ensuing, two studies are presented involving interactions between humans and computers on the interpretability of middle-to-end learning which they consider to be an important research topic for the future. While all other directions consider pre-trained networks, the authors also bring up building explainable models where methods are explored that are not a black-box approach but rather have clear semantics innately. Diagnosis of CNN representations is subdivided into five separate topics. There is a section on inspecting CNN features from an overall perspective and one on assessing areas prone to changing the output with minimal perturbation. Besides how to improve network representations

by looking at the feature space, they also discuss ways of how to detect possible biases in the CNN representation. And finally, they talk about identifying areas in the image that have a big influence on the prediction with works such as [Ribeiro et al., 2016] and [Selvaraju et al., 2017], both of which will be discussed in detail in Section 2.4 and Section 2.5 respectively. Ultimately, visualizations of CNN representations is what they consider to be the most straightforward approach to inspect latent patterns in layers. Among the mentioned works is [Zeiler and Fergus, 2013] which we will discuss in detail in Section 2.5.

2.4 Explaining Predictions

The motivation behind [Ribeiro et al., 2016] is explained with trust since the lack thereof will lead to users dismissing models or predictions. Gaining insights and understanding reasons are quite important for this aspect. Thus, the authors introduce LIME (Local Interpretable Model-Agnostic Explanations), an approach to explaining the results of any classifier. The name stems from the identified properties an explainer should have: making sense to humans and avoiding to be model-specific. LIME creates a simpler and interpretable classifier (e.g. sparse linear models or shallow decision trees) which imitates the behavior of the black-box model locally. Small changes to the input variables and observing changes in the output allows creating an explanation in the form of a list with the contributions of the features to the prediction.

DeepLIFT (Deep Learning Important FeaTures) [Shrikumar et al., 2017] is an explanation approach specific to DNNs but in turn improves the computational efficiency when compared to LIME. Instead of approximating the model, it calculates the importance of each input neuron for a prediction by a single pass of backpropagation. A score is computed by comparing to a reference (the choice of which largely relies on pre-existing domain knowledge) and reveals important parts of the input.

2.5 Visualizing Explanations for DNN Image Classifiers

The authors of [Zeiler and Fergus, 2013] refer to ImageNet [Krizhevsky et al., 2012] to make mention of the impressive performance of CNNs for image classification. They follow it up by pointing out that neural nets still lack the transparency and interpretability needed to understand their capabilities and how to efficiently improve them. To alleviate these problems, they introduce a visualization technique using a deconvolutional network (deconvnet) [Zeiler et al., 2011] in order to project feature activations back into the input pixel space which results in a feature map the size of the original image where the area (i.e. pattern) is highlighted that strongly activates the neuron. Even though this technique requires a change to the architecture of the original CNN by attaching a deconvnet to each layer, the experiments showed that their error rates lie within 0.1% of unaltered nets.

The authors [Zhou et al., 2015] criticize that the abovementioned technique using a deconvnet only works on the convolutional but ignores the fully-connected layers. They present a method to produce Class Activation Maps (CAM), a heat map to visualize where the neural net is 'looking' to identify a category. In order to obtain these discriminative regions of an image, take a weighted sum of the last convolutional layer's feature maps and then up-sample the results to the size of the input image. This goes to show that CNNs trained for classification with image labels have impressive capabilities to localize objects. Furthermore, the experiments suggest also good performances for other applications including pattern discovery, text detection, and visual question answering.

The paper [Selvaraju et al., 2017] extends the concept of CAM to Grad-CAM (Gradient-weighted Class Activation Map). The enhancements lie in the formula for calculating the CAM: as the name says the weights in the summation are the gradient (of a particular class) flowing into the last convolutional layer. It is a generalization enabling the usage of any CNN-based architecture and rids the need for both pre-training as well as changes in the network structure. The results are high-resolution representations of which parts the neural net considers important for classifying. In experiments, they show that the visualizations can help humans differentiate categories better, detect biases in data, and assess the trustworthiness of a classifier.

Interestingly enough, the authors also take time to make a point as to why transparency and explanations (and trust) are needed when interacting with AI. The progress of AI can be divided into three phases: weaker, on-par, stronger. While AI is weaker than humans at the assigned task (and therefore not reliable), transparency and explanations help to find failure modes to steer research in the right direction. As AI becomes better at what it is doing, it can be considered more reliably 'deployable'. During this stage, transparency is necessary to induce trust in users. When AI has surpassed humans, we can learn things from these systems by using explanations in machine teaching [Johns et al., 2015].

In Table 2.1 we provide an overview of the previously presented explanation techniques for explanations of predictions.

Name	**Target**	**Result**	**Technique**
LIME	Any classifier	List with weights	Approximate and simplify
DeepLIFT	DNN	Score for neurons	Calculate importance for neurons
deconvnet	CNN	Visualization	Project feature activations
CAM	CNN	Visualization	Weighted sum of last conv-layer
Grad-CAM	CNN	Visualization	Grad-weighted sum of last conv-layer

Table 2.1: Overview of the presented explanation techniques

2.6 Algorithms for the Implementation

In this final section, we discuss the two techniques used in the implementation. First, we present VGGNet, a CNN for image recognition (and classification), which was used for the labeling part of the system. To conclude, we have a look at Grad-CAM++, a method to explain predictions of a CNN visually.

2.6.1 Classification: VGGNet

The authors of [Simonyan and Zisserman, 2014] used their findings to secure the runner-up spot for the classification task in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky et al., 2015] in 2014[1]. They focus on the depth of the CNN and keep other parameters fixed. Using a small convolutional filter (3×3) and stride 1, they show that significant improvements can be achieved by pushing the number of weight layers up to 19. The net takes a 224×224 RGB-image as input and consists of a stack (of varying size depending on the chosen depth) of convolutional layers ensued by three fully connected layers with the third one carrying out the ILSVRC 1000-way classification. Soft-max is the terminal layer.

CNNs with this type of architecture were trained and evaluated on the ILSVRC 2012 data set. For the training, more than one million images were used on a system with four high-end GPUs taking two to three weeks per network. In the evaluation, two types of errors were considered: top-1 and top-5. The former describes the percentage of incorrectly classified images, while the latter covers the cases when the ground-truth is not in the top five predictions. The authors then show that their architecture outperforms all submissions of the ILSVRC 2012 on these measurements. This leads them to conclude that depth is an important aspect in CNN architecture design.

Not only does the VGGNet achieve state-of-the-art accuracy for the ILSVRC, but it also generalizes well to other tasks or data sets. Furthermore, the authors also released the weights for immediate deployment of the pre-trained networks with 16 and 19 weight layers[2].

2.6.2 Explanation: Grad-CAM++

For the explanation of the label, we chose Grad-CAM++ [Chattopadhay et al., 2018] which further improves the previously presented method Grad-CAM. The authors point out an important weakness: localizing several instances of the same class in an image leads to a decline in performance. Also, for single-object images the target might not be caught in its entirety. These two factors both affect trust negatively. In order to address these issues, they adjust the formula of Grad-CAM (see Equation 2.1) and arrive at Equation 2.2 for the weights (w).

[1] *http://image-net.org/challenges/LSVRC/2014/*
[2] *http://www.robots.ox.ac.uk/~vgg/research/very_deep/*

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \tag{2.1}$$

The improvements are achieved by adding weights (α) to the pixels to avoid a simple average and by using the Rectified Linear Unit ($relu$) activation function. The reason for the latter is explained with favoring positive influence over negative inhibition. Despite these changes, Grad-CAM++ retains the same computational complexity as its predecessor; i.e. a single backpass.

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \times relu(\frac{\partial Y^c}{\partial A_{ij}^k}) \tag{2.2}$$

The authors also introduce three metrics with all of them being expressed in percentages and computed over the whole data set: average drop, increase in confidence, and win. Average drop % is measured as the reduction in confidence when only the explanation map is used as an input instead of the full image; the lower the better the explanation. % increase in confidence is the contrary measurement to the first one and counts the number of times a boost in confidence was recorded when using the explanation map as input; since this is complementary to average drop %, higher means better. Win % directly compares the performance relying on the previous two metrics by counting the number of times one model has a higher (or lower) drop in confidence. The authors then show that Grad-CAM++ outperforms Grad-CAM on all three.

They also evaluate human interpretability of the explanations and set this criterion equivalent to invoking greater trust in users. In the experiment, they have human subjects choose the explanation map that they feel best describes the object in the image with one coming from Grad-CAM and one from Grad-CAM++ (without knowing which is which and also having the option of 'same'). Since Grad-CAM++ outperforms its predecessor significantly, the authors conclude that it instills greater trust in the underlying model.

3

Experimental Design

This chapter covers the design of the experiment and is divided into two sections. First, we recapitulate the initial research question and break it down into smaller problems. Afterward, we discuss the design of the survey we intend to use to answer the stated questions in two parts: one dedicated to the components and setup followed by one to show the structure.

3.1 Goals

This work aims to investigate trust in the context of humans interacting with AI and how knowledge of the source influences it. Inspired by the Turing Test [Turing, 1950] we set out to answer the following research question:

(RQ) Given an image classification and visual explanation of the classification, is the user's trust in the system influenced by knowing whether the answer was given by a human or by an AI?

Based on it, we can now formulate the questions we want to answer with the experiment.

(RQa) Does knowing the source influence trust?

While this might sound oddly familiar to the original research question, there is a fundamental difference. The initial statement concerns knowledge about the type of the source, whereas this case compares the availability of information against lack thereof.

(RQb) Is a machine more or less trusted than a human?

This question aims for a direct comparison of AI and human when it comes to trust. By giving the true source for a prediction and measuring trust we can try to draw some conclusions.

(RQc) Is there a bias in which source to trust?

The third and final question concerns a possible bias people might have to prefer human over AI predictions. It can be evaluated by having AI predictions disguised as produced by humans and the other way around.

With these three sub-questions in mind, we decided to conduct the experiment in the form of a survey. A priori, we knew that we could evaluate these issues with image classification and corresponding explanations. There are numerous off-the-shelf-tools available to produce both these items for a wide range of images.

By opting for a simple question-answer-survey with predefined data we ensured that we did not limit our target audience from the start. Instead of having to sit down face-to-face with participants, we also had the possibility to distribute it to a wide network of people: crowd-sourcing on the internet.

3.2 Survey Design

In order to find answers to the questions introduced in the previous section, we designed a survey. First, we present the components and then the structure we arranged them in.

3.2.1 Components and Setup

Since all of our research questions are about comparing trust in different settings, we needed a measure that quantifies it and allows us to compare. We found such a measure in the Multi-Dimensional Measure of Trust (MDMT) [Ullman and Malle, 2019]. The authors propose 16 items to be evaluated on a discrete scale from 0 (not at all) to 7 (very) and including an option for 'not applicable'. They are grouped into four dimensions (reliable, capable, ethical, sincere) of four items each. Furthermore, the authors identify two factors of trust: capacity (reliable, capable) and moral (ethical, sincere). As the name says, the MDMT is a measure of trust in the context of human-machine (or human-human) interaction. This allowed us to calculate several different measures for trust by averaging the values for a dimension or a factor. For a full list of the items as well as the evaluation scale, please refer to the original publication [Ullman and Malle, 2019].

Another questionnaire we integrated into the survey is the Affinity for Technology Interaction (ATI) scale [Franke et al., 2019] consisting of nine questions. It gives statements and asks if the participants agree or disagree on a six-point Likert scale. For a full list of the associated questions, please refer to the appendix of the original publication [Franke et al., 2019]. The authors report studies to have shown "[...] moderate to high correlation with geekism [and] technology enthusiasm [...]". We expect these two things to be connected to acceptance and trust for a new system.

Lastly, we also want to learn about the participants' stance regarding trust in general. For this purpose, we used the three questions of the SOEP-trust (Socie-Economic Panel) survey proposed in [Naef and Schupp, 2009]. They all consist of a statement and a four-point Likert scale on agreeance. With these questions, we gained insight into the attitude of the participant when it comes to trusting strangers.

Regarding the setup, we opted to assign people to one of nine groups. They all represent a different setting to compare trust in. Table 3.1 provides an overview. We can split them into two categories: those who get an explanation (groups 1, 2, 3) and those who do not (groups 4 to 9). Everyone sees the original images that were used as input for the classification algorithm along with the produced label. Three groups (3, 8, 9) are not given a source for label or explanation, denoted by '?' in the row 'Given'. Instead, they are told that these two things were produced by an unknown source (but limited to human or AI). Groups 1 and 2 truthfully get the label produced by human (H) and AI respectively. The same goes for groups 4 and 6 which additionally also see an explanation for the label. Groups 5 and 7 are being deceived and see different explanations than they are being told, i.e. human-produced for AI and AI-produced for given source human. We aimed to show the tendencies people have to trust or distrust an entity.

Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Label	YES	YES	YES	YES	YES	YES	YES	YES	YES
Explanation	NO	NO	NO	H	H	AI	AI	H	AI
Given	H	AI	?	H	AI	AI	H	?	?

Table 3.1: Overview of the different groups for the experiment

3.2.2 Structure

A rough overview of the designed survey can be seen in Figure 3.1. It consists of four parts in three blocks. The first part serves as the introduction and includes some initial questions. It is followed by the main block containing six images (as well as explanations where applicable) and the corresponding questions. Part three is also still in the main block and consists of a questionnaire about the set of all six images. The last and third block is made up of three final questions regarding general trust in strangers.

The first part of the survey is prefaced by a brief overview of what the study is about: exploring the relationship between humans and machines as well as comparing it to human-human interaction. After asking for the participant's informed consent, we gather some basic demographic traits such as age group, country, etc. We also utilize an attention task in this first part. We pose the question about the experience with the crowd-sourcing platform but instruct them in a short paragraph to ignore said question and put something else in the answer box. This allows us to discard answers for which we have to assume that the participant did not read the instructions carefully and thus

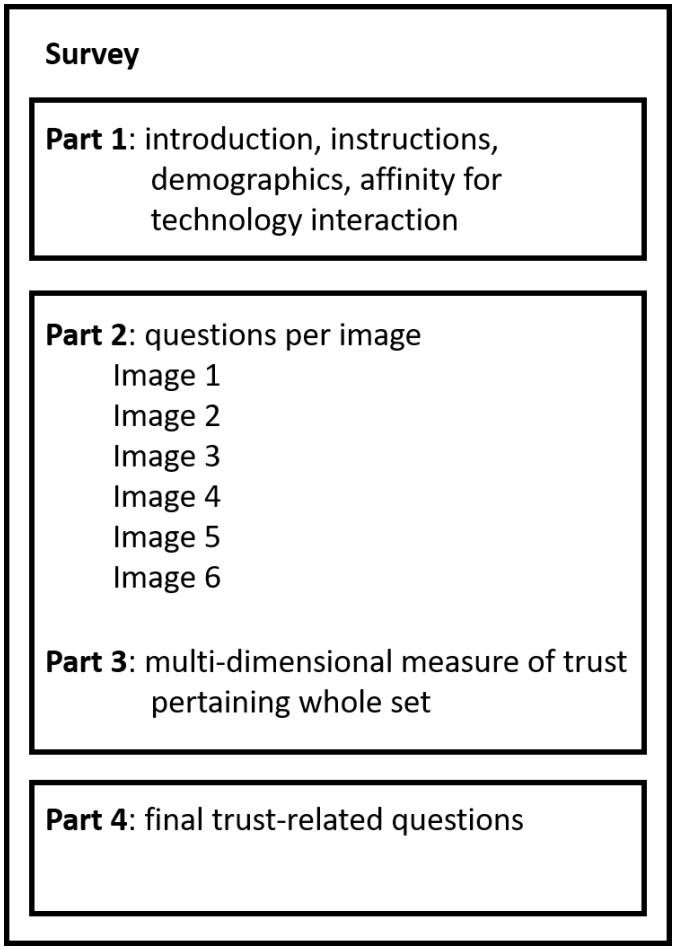


Figure 3.1: Structure of the survey

renders their replies unusable.

The second half of the first part consists of the ATI scale to gather data on the technology affinity of the participants.

The second part starts with a short recapitulation of the instructions along with a declaration of the given source. Then follows the centerpiece of the survey: a set of six images and their labels. The label consists of one or multiple keywords that describe an object present in the image. Depending on the group, they are shown along with an explanation, a visualization of the reasoning for the label. For each image, we ask the participant if the label is correct and in case they selected no have them elaborate and give a new label. Where applicable, we do the same for the explanation: does it capture the label well? We also ask for a reasoning if they deny as well as their proposed improvement for the explanation (move, resize, or both). Those things are done for each of the six images individually.

We show multiple images from the same source to give the participants something to build an opinion on. One of the six images has a fake and obviously wrong label in order to keep accuracy constant and to ensure that it is not a confounding factor. Introducing an error involves risk and therefore trust.

The third part shows all the original images (without label or explanation) and asks the questions of the MDMT. We use it to evaluate the trust in the (given) source.

We close the survey out with the SOEP-trust-questionnaire. This allows a comparison between trust in strangers and trust in the explicitly mentioned entity from the main block.

For an impression of what the final survey looked like, please refer to Appendix B for screenshots.

4

Implementation

This chapter covers everything we implemented. We start by describing the framework built with off-the-shelf-tools and used to produce the data necessary for the experiment proposed in Chapter 3. Next, we elaborate on the data set and its subset we chose as input. We then discuss our efforts to make AI and human explanations comparable. This included bringing them to a common format and their visualizations. Finally, we talk about the implementation of the experiment using the crowd-sourcing platform Amazon Mechanical Turk.

The code for the implementations can be found on the accompanying CD. For detailed instructions on how to use the framework, please consult Appendix A.

4.1 Framework and Modes

We created a framework that can be used to produce explanations for a multitude of input images. The user simply needs to provide said data along with some code for classification and explanation. The combination of these two techniques into a single file is called *mode* in this project. The first such mode was built based on the code for the implementation of Grad-CAM++[1] provided by the paper [Chattopadhay et al., 2018]. It combines their proposed explanation technique with VGGNet [Simonyan and Zisserman, 2014] and works out of the box with any input image. We started by porting the code to Python 3 and removing unnecessary outputs. Next, we rearranged the code to be better comprehensible and to make it easier to build similar modules. This included strictly sectioning the code into parts for classification and explanation.

The implementation for the classification using the deep convolutional neural network is done with TensorFlow [Abadi et al., 2016]. Pre-trained networks have been made publicly available[2] and this allowed us to skip the time-consuming task of training it ourselves. The implementation uses one with 16 weighted layers (hence called VGG16). In the technical report [Simonyan and Zisserman, 2014] it is referred to as ConvNet

[1] *https://github.com/adityac94/Grad_CAM_plus_plus*
[2] *http://www.robots.ox.ac.uk/~vgg/research/very_deep/*

Configuration D. It was trained using the training set of the ImageNet Challenge 2014[3], which consisted of almost half a million images and contained 1000 classes.

The code for the classification of the images loads said neural network as well as the input image. After the execution, the label with the highest predicted probability of fitting the image is extracted. We added a utility file with some helper functions. It contains the code to load, crop (to have the required 224×224×3 dimensions), and save the input image for the neural net as well as the probing to only get the top predicted label from the neural network since we only rely on one in this scenario.

Furthermore, the utility file also visualizes the explanation produced by the AI. But before that, it has to be calculated first. Also using TensorFlow, the central Grad-CAM++ Equation 2.2 is implemented with the Python library NumPy [Oliphant, 2006]. The resulting CAM is resized to fit the dimensions of the input image and passed to the function responsible for the visualization.

4.2 Data Set

The data set was chosen according to several requirements: high-resolution images, human-produced (and not just human-verified and machine-produced) labels, a broad selection of classes (instead of only covering one special field), and ideally also providing some sort of explanation. Only one of the available data sets was identified to fulfill all of these criteria: ImageNet [Russakovsky et al., 2015]. It contains more than 14 million high-resolution images in nearly 22'000 categories. A subset of over one million images also have bounding box annotations[4], which we can use as explanations. The images were collected on image hosting services and search engines like Flickr[5]. Annotations (labels as well as bounding boxes) are human-made and human-verified using crowd-sourcing.

4.3 Data Selection

The data set for the survey was built incrementally using the framework and example mode described in Section 4.1. We set the goal to have 60 images labeled and explained by both human and AI. This resulted in a data set of 180 images; one-third of which were the cropped original images and one third each human and AI explanations.

As input we used the validation data set of the ImageNet Large Scale Visual Recognition Challenge 2011 (ILSVRC2011)[6] [Russakovsky et al., 2015]. One might argue that it is not methodologically correct to use a set that was somehow involved in training. We counter this argument with the fact that the test set does not include the properties of the bounding boxes which are crucial for our use case. Since we try to avoid the training

[3] *http://www.image-net.org/challenges/LSVRC/2014/*
[4] *http://image-net.org/about-stats*
[5] *https://www.flickr.com/*
[6] *http://www.image-net.org/challenges/LSVRC/2011/registered-downloads*

set at all costs, we were left with the validation set. Also, we were more interested in AI explanations than labeling accuracy.

With a Python script (available in the repository) we randomly chose images from the 50'000 available while keeping track as not to choose two from the same class. We then ran the framework on these data batches and discarded elements according to a number of guidelines. First of all, the label given by the AI had to coincide with the human label to avoid any possible bias from errors. Furthermore, the image could not depict anything that we considered to be inappropriate content (e.g. a possibly dead animal). We also rejected images that had multiple instances of an object in order to balance the scales between human and AI detection (and therefore explanation). According to their paper [Chattopadhay et al., 2018], this is a known weakness of their predecessor's approach and our results confirmed that flaw to still be existing in Grad-CAM++ to a certain extent. Comparing this factor is considered out of scope for this work. Lastly, we checked the images to have labels that make sense for a layman as not to inadvertently have subjects flag images as wrongly labeled because the depicted concept is too complicated. We repeated this process until we had the desired data set of 60 images completed.

Finally, we chose ten out of the 60 images at random and gave them an obviously fake label (for both human and AI) in order to introduce some inaccuracy into the results, while keeping the explanations. This resulted in about 83% correctly labeled images for the final data set.

All the produced images are available at *https://files.ifi.uzh.ch/MTxai19/*.

4.4 Toward Comparable Explanations

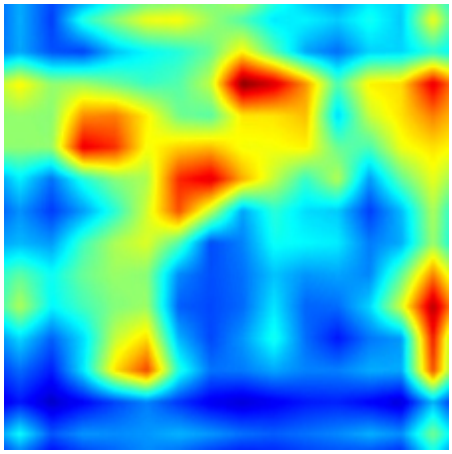
[Zhou et al., 2015] point out that their explanations can be used for localization. While the bounding boxes of ImageNet mainly act as localization, they can be seen as part of an explanation. To the best of our knowledge, the explicit connection of localization to explanation remains to be shown and is left up to future work.

The initial problem was that we started from two different formats: the human explanation was a bounding box given by the coordinates of the four corners and the AI explanation was a heat map (see Figure 4.1a). For the experiment, we needed them in a state where they are comparable.

Human and AI explanations are both produced by the previously mentioned utility file (created specifically per mode). In our example, two functions are responsible: the one loading the image also creates the human explanation, while there was one dedicated to visualizing the output of the Grad-CAM++-algorithm. We will now discuss the process of bringing the two explanations into a comparable format as well as the visualization methods.

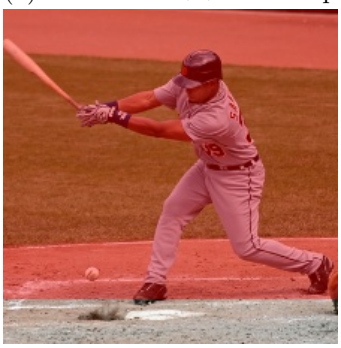
4.4.1 Converting the Heat Map

The result of the Grad-CAM++-algorithm is a 224×224 matrix with values ranging from 0 to 1, denoting the impact of a pixel on the label. In the default implementation,



(a) Grad-CAM++ heat map

(b) Threshold: 10%



(c) Threshold: 50%

(d) Threshold: 75%

(e) Threshold: 85%

(f) Threshold: 90%

(g) Threshold: 95%

Figure 4.1: Different threshold values for the conversion from heat map to bounding box

the visualization of the explanation is a heat map with red (important), green (middle-ground), and blue (unimportant) areas.

We were now presented with four options to make the explanations comparable: convert the heat maps to bounding boxes, the bounding boxes to heat maps, both to a new common format, or produce new human explanations. We discarded the last option

since it was clearly out of scope considering the given frame, time, and resources. The third choice was also beyond the intended extent of this work since it would have made investigations necessary as to which format would be suitable or ideal. This left us with the conversions. In order to produce a heat map from a bounding box, the introduction of some noise would have been necessary to not only color the areas but to also frazzle the boundaries. In this case, the human explanation could no longer be considered human-produced since it would have been machine-processed (we already covered involving humans in creating new explanations above). The only feasible option was to convert the heat map to a bounding box.

Luckily, there were tools available for such a task and we only needed to define the parameters. Using the Python library OpenCV's [Bradski, 2000] thresholding we converted the heat map to a binary (black and white) image. The resulting contours were then processed with the function *boundingRect* to compute a bounding box.

The threshold for the conversion is defined as a parameter in the utils-file as a percentage value (i.e. between 0 and 1) and can be adjusted. We decided on the value used in our experiment empirically. Figure 4.1 shows the initial Grad-CAM++ heat map for an example image as well as the bounding boxes for several thresholds in the form of colored areas. We computed multiple AI bounding box for a set of images and compared them. The main goal was to capture all of the 'bright red' regions in the heat map and to exclude unimportant areas. Furthermore, we also tried to avoid making it obvious which explanation came from which source as this would lead to bias in the experiment. We finally settled on the value of 75% as it would produce the most consistent results over the range of images we tested.

4.4.2 Visualizing the Explanation

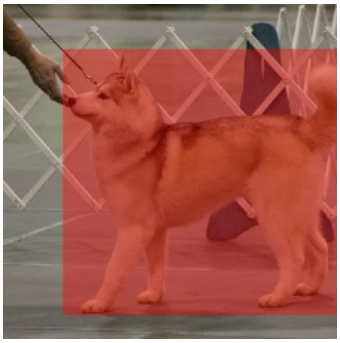
Now with both explanations in the same format (i.e. given by corners of a bounding box), we needed to decide on the visualization of the explanations. We produced them in three different ways, as seen in Figure 4.2, for a series of images.

Figures 4.2b and 4.2e show our initial idea: the bounding box is represented as a red area overlay on the original image. While it might seem better to only show the outlines of the rectangle, we decided against it since we could not be sure that the chosen color would easily be visible against the overall hue of the image. Therefore, we went with a filled area to alleviate this problem. Furthermore, we had a variable in the utils-file that controlled the alpha of the overlay. We empirically determined 0.4 to be a good value that would make both the bounding box as well as the underlying original image visible. Both Figures 4.1 and 4.2 are produced using this value.

We also considered the inverse, as seen in Figures 4.2c and 4.2f. Instead of placing the emphasis on a certain area by coloring it in, we did it the other way around. Everything except for the determined bounding box was filled with the color white. This left the area that was considered important in place but removed its surrounding context.

Figures 4.2d and 4.2g show the variant which cuts the area out and displays it separately. It is then scaled to about the size of the input image. In addition to removing

(a) Normalized input image



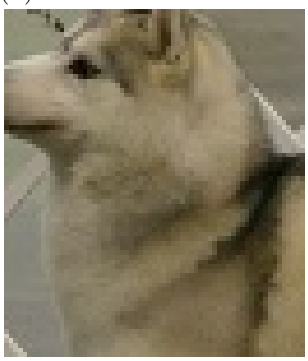
(b) Human: colored area

(c) Human: cut out in place

(d) Human: cut out scaled

(e) AI: colored area

(f) AI: cut out in place



(g) AI: cut out scaled

Figure 4.2: Different visualizations for a bounding box explanation

the context, it also takes out the positional information of the bounding box as well as the proportions to the image.

In the end, we decided to use the colored area since we deemed the context of the bounding box too important. Seeing only a part of the image could be confusing in some cases and even more so when they were scaled differently. We kept the color of the area

red and manually checked the produced images and explanations to make sense and be understandable with it.

4.5 Deployment on Amazon Mechanical Turk

After obtaining all the necessary data (images, labels, explanations), we decided on using the crowd-sourcing platform Amazon Mechanical Turk (AMT) for the survey. Our requirements coincided exactly with the description of a Human Intelligence Task (HIT)[7], a single self-contained job that can be completed by a worker. In our case, a HIT consists of one run-through of the survey.

AMT allowed us to create a template in HTML (Hypertext Markup Language) and read data from an additional Comma-Separated Values (CSV) file. Once uploaded and deployed as a project, AMT took care of connecting it to the workers, gathering their answers, and also producing a CSV that we could use for the evaluation.

The final implementation of the survey was composed of two parts: the HTML being responsible for displaying the questionnaire as well as containing some logic for handling the input from the CSV and a simple web Application Programming Interface (API) to ensure workers could only participate once since AMT did not provide such a functionality.

The HTML mainly consists of the questions and containers for the images and labels. While they are empty by default, JavaScript was used to fill them with the corresponding contents. Through the functionality of AMT we can provide a CSV with 90 rows of which every single one holds all the necessary data to represent one HIT. We distributed the 60 original images to ten buckets of six images each while making sure that each bucket has one image with a fake label. Each line holds the group number, the source given to the participant as well as the original image plus explanation (where applicable). Nine groups with ten buckets each resulted in the above mentioned 90 rows. With placeholders in the variable definitions, the data is automatically read from the CSV. We used JQuery[8] for the management of the elements. Not only are they filled this way but we can also hide and show what we need. The necessary logic is implemented in the ready-function which is executed as soon as the site loads. The survey form is hidden by default and only shown after a query to the API to ensure that workers only participate once. This call is done with Ajax (Asynchronous JavaScript and XML) and the form is unlocked in the callback-function, provided that it is the worker's first access to the HIT. During this procedure, we also set all answer fields to empty to track which questions were answered and which were ignored. We then hook into the on-submit-function of the form and add another call to the API to register the worker's ID as a participant. Next, we randomize the order of the six images read from the CSV and put them in the predefined containers. Finally, we show and hide some elements pertaining to source and explanation in order to display the correct information for the worker's group.

[7] *https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/mechanical-turk-concepts.html*
[8] *https://jquery.com/*

To keep track of the participants and to prevent them from completing more than one iteration of the survey, we used the micro web framework Flask[9]. We deployed the instance on a free account of *https://www.pythonanywhere.com* since it provided all the functionality we needed. The API has two entry points: one to add and one to query workers. As soon as they submitted their answers, we created a save file for that worker ID and included the group number for possible later use. The other method is called on load in the HTML. It queries the server if there already exists a save file for a certain worker ID and receives the respective state. That way, we could determine if the worker had already answered the survey and hide it to prevent them from completing it again.

This construct was then deployed on AMT as a survey and accessed by 900 unique workers in order to gather their responses. AMT provided the results in a downloadable CSV including some additional data like average and individual response time.

[9] *http://flask.pocoo.org*

5

Evaluation

This chapter discusses the findings from the results obtained in the experiment described in Chapter 3. After deploying it to Amazon Mechanical Turk and having 900 unique workers take the survey, we analyze their responses. Both the data set as well as the code used in the evaluation are available on the accompanying CD.

To start, we recapitulate the questions we want to answer beginning with the initial research question followed by the three problem statements as described in Section 3.1.

(RQ) Given an image classification and a visual explanation of the classification, is the user's trust in the system influenced by knowing whether the answer was given by a human or by an AI?

(RQa) Does knowing the source influence trust?

(RQb) Is a machine more or less trusted than a human?

(RQc) Is there a bias in which source to trust?

The survey on AMT resulted in a CSV that could easily be loaded using Python which we used for the evaluation. Every submission by a participant corresponded to one line, so we had 900 data rows in the CSV. In a first step, we reduced the number of columns from 438 to the 20 relevant for our questions. This included the MDMT [Ullman and Malle, 2019] and some demographics. We then filtered out participants who did not give consent to use their information (0% by design with 100% retained). Subsequently, we removed answers where the attention task had been failed. We were more tolerant in this check than we initially intended to be. Answering the given question rather than the attention test or ignoring it resulted in rejection. But instead of insisting on pinpoint accuracy, we also accepted variations of the expected answer due to typing mistakes (such as lower and upper case) or slightly misreading (e.g. answering with two sentences instead of two words).

Still, it resulted in 445 rows being removed (49% and therefore 51% retained). Out of the original 100 participants per group we were left with groups of sizes between 41 and 61 (see Figure 5.1). We deemed it enough to get relevant results. The remaining 455 rows were preprocessed for the MDMT: we replaced the corresponding number

for 'not applicable' (8) with 'not a number' so they would be ignored for the ensuing computations.

To answer the questions, we calculated trust as measured by the MDMT in both the four dimensions (reliable, capable, ethical, sincere) as well as the two factors (capacity, moral) by averaging the corresponding items. The MDMT assigns values from 0 (not at all trusted) to 7 (very trusted). We compared these elements for previously defined selections of groups and combinations thereof. An overview of the experiment groups can be seen in Table 3.1. We address RQa, RQb, and RQc before approaching RQ.

In order to check if the data at hand was normally distributed, we used SciPy's normaltest[1]. We report all the obtained p-values for every sample's trust dimensions and factors in Appendix C. For most combinations, we received p-values significantly smaller than 0.05. This suggested that the majority of the samples was not normally distributed.

With this knowledge, we decided on non-parametric tests. We used the two-sided Kolmogorov-Smirnov statistic[2] for the comparison of two sets of groups. The Kruskal-Wallis H-test[3] was used for three or more sets. Where necessary, we opted for Dunn's test[4] as post-hoc pairwise comparison. We report all the numbers in the following tables rounded to three decimal places.

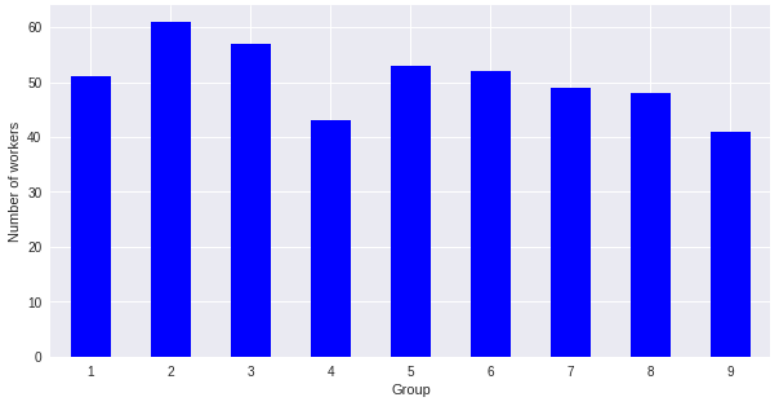


Figure 5.1: Number of workers per group remaining after attention task test

Caveat! While preparing the data for evaluation, we noticed that we had accidentally dropped one item of the MDMT for the survey, namely respectable. We replaced the corresponding column in the data with 'not applicable' (i.e. 'not a number') and calculated the respective trust dimension (ethical) and factor (moral) nevertheless. 'Not applicable' is an answer option defined by the MDMT and according to the authors the measurements computed despite missing values still hold.

[1] *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html*
[2] *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html*
[3] *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html*
[4] *https://github.com/maximtrp/scikit-posthocs*

(RQa) Does knowing the source influence trust?

To find an answer to this question, we combined groups 1 and 2 into a set of participants that had received the information about the source which had produced the labels. We compared their trust measurements against those of group 3 which had not been given the source. Table 5.1 shows the trust values as measured by the MDMT. While the difference column implies that the participants rated the known source higher than the unknown source, the p-values obtained from the Kolmogorov-Smirnov statistic indicate that none of the differences are statistically significant.

Dimension / Factor	**Known source**	**Unknown source**	**Difference**	**p-value**
Reliable	4.901	4.823	0.078	0.749
Capable	4.991	4.880	0.111	0.763
Ethical	4.989	4.791	0.199	0.174
Sincere	5.030	4.884	0.147	0.552
Capacity	4.946	4.852	0.094	0.667
Moral	5.043	4.848	0.195	0.260

Table 5.1: Trust as measured by the MDMT for RQa: known (groups 1 and 2) compared to unknown source (group 3) with no further information

Table 5.2 shows the measured trust in known and unknown sources when adding explanations. We compared the combination of groups 4 and 6 to the union of groups 8 and 9. Even though the unknown source exhibits a greater value on both the capable and sincere subscale, none of the differences are statistically significant.

Dimension / Factor	**Known source**	**Unknown source**	**Difference**	**p-value**
Reliable	4.632	4.566	0.066	0.308
Capable	4.624	4.627	-0.004	0.589
Ethical	4.692	4.662	0.029	0.850
Sincere	4.823	4.840	-0.017	0.937
Capacity	4.628	4.596	0.031	0.180
Moral	4.787	4.757	0.030	0.875

Table 5.2: Trust as measured by the MDMT for RQa: known (groups 4 and 6) compared to unknown source (groups 8 and 9) with explanations

(RQb) Is a machine more or less trusted than a human?

Table 5.3 contains statistically significant differences (denoted by * next to the p-value). Since we used a two-sided test, we can consider results with both $p < 0.025$ as well as $p > 0.975$ indicative. We compared trust in labels and explanations produced by a human (group 4) to AI-made (group 6). Those groups were given the actual source type. The capable subscale and the capacity trust both exhibit statistically significant differences. According to their answers, participants trust humans more in the capable dimension but slightly prefer AI for the capacity factor.

Dimension / Factor	Human	AI	Difference	p-value
Reliable	4.616	4.644	-0.028	0.708
Capable	**4.634**	**4.615**	**0.018**	0.993 *
Ethical	5.004	4.409	0.596	0.087
Sincere	5.043	4.631	0.411	0.499
Capacity	**4.625**	**4.630**	**-0.005**	0.998 *
Moral	5.044	4.563	0.481	0.106

Table 5.3: Trust as measured by the MDMT for RQb: labels and explanations from a human (group 4) compared to those from an AI (group 6)

In the second step toward answering this question, we created two samples by combining two groups each. Groups 4 and 7 had both been given human as the source but the latter had seen explanations and labels from an AI. The other sample consisted of groups 5 and 6 who were told an AI was the source but in truth for the latter it was actually a human. As seen in Table 5.3, participants showed more trust toward the AI as the given source with the exception of the ethical dimension. But those results are inconclusive since none of the differences are statistically significant.

Dimension / Factor	Human	AI	Difference	p-value
Reliable	4.491	4.861	-0.37	0.425
Capable	4.440	4.863	-0.423	0.081
Ethical	4.708	4.659	0.049	0.247
Sincere	4.757	4.834	-0.077	0.612
Capacity	4.465	4.862	-0.397	0.102
Moral	4.753	4.778	-0.025	0.396

Table 5.4: Trust as measured by the MDMT for RQb: given source human (groups 4 and 7) compared to given source AI (groups 5 and 6)

(RQc) Is there a bias in which source to trust?

In order to detect a possible bias, we provided the participants with fake sources. We did this for both human as well as AI labels and explanations. First, we compared groups that had seen human-generated products but only one of those had them designated truthfully (group 4). The other two were told it had been an AI (group 5) or an unknown source (group 8). Table 5.5 reports the measured trust along with the p-values. Only two rows show statistically significant differences: reliable and capacity.

Dimension / Factor	**H as H**	**H as AI**	**H as ?**	**p-value**
Reliable	**4.616**	**5.074**	**4.502**	0.038 *
Capable	4.634	5.105	4.615	0.071
Ethical	5.004	4.903	4.527	0.099
Sincere	5.043	5.033	4.672	0.157
Capacity	**4.625**	**5.090**	**4.558**	0.031 *
Moral	5.044	4.989	4.604	0.091

Table 5.5: Trust as measured by the MDMT for RQc: comparison of groups seeing labels and explanations made by a human but only told so truthfully once (group 4) while the other two are given AI (group 5) or unknown (group 8) as the source

Using the post-hoc test, we did a pairwise comparison of the relevant measurements. The p-values in Tables 5.6 and 5.7 indicate that the statistically significant differences are between groups 5 and 8. For both the reliable dimension as well as the capacity factor the participants trusted more in the human posing as AI than in the human disguised as an unknown source. The numerical differences found were 0.572 for reliable and 0.531 for capacity.

	H as H	**H as AI**	**H as ?**
H as H	-1	0.402	0.402
H as AI	0.402	-1	0.032 *
H as ?	0.402	0.032 *	-1

Table 5.6: p-values for the pairwise comparison of the possibly statistically significant differences for 'reliable'

	H as H	**H as AI**	**H as ?**
H as H	-1	0.198	0.401
H as AI	0.198	-1	0.029 *
H as ?	0.401	0.029 *	-1

Table 5.7: p-values for the pairwise comparison of the possibly statistically significant differences for 'capacity'

We did the same for AI-generated explanations and labels: comparing samples given the true source (group 6), disguised as created by a human (group 7), and posing as unknown source (group 9). The three-way test as reported in Table 5.8 did not show any statistically significant differences between the samples.

Dimension / Factor	**AI as AI**	**AI as H**	**AI as ?**	**p-value**
Reliable	4.644	4.378	4.640	0.755
Capable	4.615	4.266	4.642	0.360
Ethical	4.409	4.468	4.833	0.336
Sincere	4.631	4.514	5.048	0.213
Capacity	4.630	4.322	4.641	0.427
Moral	4.563	4.504	4.947	0.269

Table 5.8: Trust as measured by the MDMT for RQc: comparison of groups seeing labels and explanations made by an AI but only told so truthfully once (group 6) while the other two are given human (group 7) or unknown (group 9) as the source

(RQ) Given an image classification and a visual explanation of the classification, is the user's trust in the system influenced by knowing whether the answer was given by a human or by an AI?

To answer the initial research question, we summarize the results of the sub-questions. The findings of RQa suggested that there is no difference in trust from knowing or not knowing the source. This was confirmed by the second part of RQc. Even though we disguised AI products as human and unknown source alongside revealing the true nature, participants did not exhibit a clear preference. This was indicated by the absence of a statistically significant difference and implies that knowing the source does not influence trust.

The first part of RQc (human explanations and labels given as true and fake sources) may suggest otherwise and also show statistically significant differences but we consider these results inconclusive. A clear preference was exhibited toward human source disguised as AI over human as unknown for the reliable dimension and the capacity factor. It is not evident, however, whether this implies increased trust in the human explanation or the fake AI source as the measurement gives no indication for that. Additional examination of the existing data is required. This could possibly include a comparison of the ATI values for the participants which is considered out of scope for now.

The evaluation of RQb also yielded no decisive results regarding a potential preference. Participants exhibited minimally higher trust for both human (capable dimension) and

AI (capacity factor) explanations and labels. Since every other difference was statistically not significant, we have no clear proof for a possible preference.

To conclude and give an answer to the research question, we can say that we did not find any heavy indications for trust to be influenced by the availability or information itself about the source of an image classification and its visual explanation in a human-machine-scenario.

6

Limitations and Future Work

This chapter covers the limitations we identified and explains how they can be addressed in future work. It is divided into two sections. First, we describe the restrictions regarding explanations. We then point out additional work to be done in the future.

6.1 Explanations

In Section 4.4, we mentioned that [Zhou et al., 2015] find their explanations can be used for localization. We left the reverse direction - are localizations an explanation? - for future work. This is out of scope here since it also begs the following question: what makes a good explanation? Not only are more experiments needed to answer this in the context at hand but it requires the evaluation of explanation techniques and their visualizations.

While we introduced several explanation approaches in Section 2.4, we limited the experiment to one. Furthermore, since there is an endless multitude of image classification algorithms and we can combine each of them with an explanation method, we end up with a sheer infinite amount of modes to be used in the framework. We suggest some evaluation beforehand to limit the numbers but then it might be interesting to compare them in a large scale experiment in a trust setting.

Instead of limiting the diversity of the different explanations to visualization like the three presented Subsection 4.4.1, one could explore different possibilities. While comparing different visualizations (heat map, bounding box, etc.) might certainly be interesting, the explanations do not have to be limited to a projection back to the image space. Of the introduced techniques in Section 2.4 some use a list of features for explanations. This concept can be extended to keywords or even generalized to full sentences giving a description. More investigation is required.

We declared producing new human explanations out of scope in Subsection 4.4.1. With existing technologies such as eye trackers human subjects could explain the label of images by generating heat maps of the areas they focus on. These could then be directly compared to the heat maps generated by Grad-CAM++ [Chattopadhay et al., 2018] and the likes.

6.2 Future Work

In our experiment, we intentionally limited the data used to 60 images in order to be able to address more people and therefore get more responses. For the next step, one should think about repeating the experiment with more or different data. It could also be interesting to let people use their own images and have them rate the trust generated by the explanation. A similar setup (without rating) is already implemented by Grad-CAM[1] [Selvaraju et al., 2017].

In the current setup, we imposed the restriction for each image to have the same label, irrespective of the source. In other words, the labels produced by AI and human for an image used in the experiment were the same. It might be an interesting approach to remove the constraint that AI and human have to coincide with the label. This allows to study the impact of accuracy and see which prediction generates better trust also with respect to the explanation. Performance is mentioned as a reason for trust in AI in [Siau and Wang, 2018]. Do people have reservations toward AI and is possibly better performance enough to cancel out their bias and trust the machine more?

Furthermore, the data used in the experiment does not have to be confined to image classification. Using different tasks as the centerpiece with the same approach allows to check if the results for trust are consistent.

Also, a different approach to the experimental design could prove to be interesting. Instead of aiming to have many different people answer a survey, extend its length and get more consistent and coherent results from fewer participants. We consider our study to be a first check to see if we even get a signal which is why we wanted to get as many different contributors as possible. Such an extended experiment might be beneficial to obtain a potentially more accurate measure of trust. It could also be used to investigate the evolution of trust over the course of the experiment as it is dynamic and can change drastically [Hoff and Bashir, 2015].

We also want to point out that we analyzed only a fraction of the available data, namely the trust as quantified by the MDMT [Ullman and Malle, 2019]. We have not even touched upon the possible relationship between trust and the measured ATI [Franke et al., 2019]. Furthermore, the data regarding the effect of faking the source and the implied deception went mostly ignored for now.

The evaluation could also be expanded in other ways, for example with improved filtering of the answers. While we employed a simple attention task, one could also check for patterns in the answers to find workers with the sole goal of finishing as quickly as possible. We tested for entries with only one unique answer in the MDMT (i.e. had the same option for all items) but decided not to exclude them as they might have been genuine answers. Further analysis is required for this. It is worth noting that we had a function in place to stop the time participants took to fill in the survey and possibly filter them on it for the evaluation. However, the way workers engage HITs on AMT rendered the measurements useless because opening a task does not imply that they start right away. An improved implementation is necessary.

[1] *http://gradcam.cloudcv.org/classification*

The experiment could also be repeated (including all of the 16 items of the MDMT) with a higher number of participants to possibly get a more accurate measurement for trust and statistically significant differences. It might also be interesting to only use master workers on AMT and see if there is an increase in the quality of the answers.

Last but not least, the MDMT as a measurement has to be challenged with regard to the results obtained which did not allow clear conclusions. One case indicated higher trust on the capable subscale toward a human but the capacity trust (which is a composite of the two dimensions reliable and capable) showed increased trust in AI over a human. We propose to either directly ask for a value of trust on a scale (which is to be determined) or at least use an additional measurement for trust to compare the MDMT to.

7

Conclusions

We investigated how the availability of certain information affects trust of humans when interacting with AI. The goal was to find an answer to the question if trust was influenced by knowing the source (human or AI) of a prediction and its reasoning. We broke it down into three sub-problems. Firstly, does knowing the source influence trust? Secondly, is a machine more or less trusted than a human? Finally, is there a bias in which source to trust?

First, we explored what trust is in the setting of human-machine-interaction and how to build it with explanations. Subsequently, we put the focus on neural network image classifiers and how to visualize their reasoning. Using the popular CNN VGGNet [Simonyan and Zisserman, 2014] for classification and Grad-CAM++ [Chattopadhay et al., 2018] to visually explain the predictions we built a framework to enable batch processing of images.

With the questions above in mind, we designed a survey incorporating images along with labels and explanations produced by different sources. In ImageNet [Russakovsky et al., 2015] we found a data set which had labels and explanations (bounding boxes for the image classification task) generated and verified by humans. The AI counterparts were provided by the above-mentioned framework. Bringing the explanations into a common format meant converting the heat maps produced by the AI into bounding boxes. This allowed us to make comparisons between the two. By having the same labels for both sources we ensured that accuracy and therefore performance did not play a role for trust.

In the survey, we showed each human subject a set of six images, labels, and the type of the source (human, AI, or unknown). Depending on which of the nine groups the participants were in they also received the visualized explanations. The MDMT [Ullman and Malle, 2019] allowed us to measure the trust the subjects had established with six different values (four dimensions and two factors). We deployed the questionnaire to 900 human workers on AMT which resulted in 100 participants per group.

While evaluating, we compared the trust measurements between previously defined selections of groups and combinations thereof. Calculating the differences between them and checking for statistical significance, we gained several insights. We could not detect an effect on trust of knowing the source or not. Similarly, no unambiguous bias was found toward any type of source. There were two statistically significant differences for

trust in machines versus trust in humans. But since they were both very small and one was in favor for AI and one for human, we did not receive a clear result.

This led to the conclusion that knowledge about the source and the availability thereof does not exhibit a distinct influence on trust of humans in a machine.

References

[Abadi et al., 2016] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:1603.04467*.

[Bradski, 2000] Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

[Chattopadhay et al., 2018] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE.

[Dosilovic et al., 2018] Dosilovic, F. K., Brcic, M., and Hlupic, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215. IEEE.

[Franke et al., 2019] Franke, T., Attig, C., and Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human-Computer Interaction*, 35(6):456–467.

[Freitas and A., 2004] Freitas, A. A. and A., A. (2004). A critical review of multi-objective optimization in data mining. *ACM SIGKDD Explorations Newsletter*, 6(2):77.

[Goodman and Flaxman, 2017] Goodman, B. and Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'. *AI Magazine*, 38(3):50.

[Gunning, 2017] Gunning, D. (2017). Explainable Artificial Intelligence (XAI).

[Hayes and Ford, 1995] Hayes, P. and Ford, K. (1995). Turing test considered harmful. In *IJCAI (1)*, pages 972–977.

[Hoff and Bashir, 2015] Hoff, K. A. and Bashir, M. (2015). Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3):407–434.

[Israelsen and Ahmed, 2019] Israelsen, B. W. and Ahmed, N. R. (2019). 'Dave...I can assure you ...that it's going to be all right ...' A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships. *ACM Computing Surveys*, 51(6):1–37.

[Johns et al., 2015] Johns, E., Mac Aodha, O., and Brostow, G. J. (2015). Becoming the Expert - Interactive Multi-Class Machine Teaching.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks.

[Langley et al., 2017] Langley, P., Meadows, B., Sridharan, M., and Choi, D. (2017). Explainable Agency for Intelligent Autonomous Systems. *Twenty-Ninth IAAI Conference*.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., and others (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Mayer et al., 1995] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Academy of Management Review*, 20(3):709–734.

[Moor, 1976] Moor, J. H. (1976). An analysis of the turing test. *Philosophical Studies*, 30(4):249–257.

[Naef and Schupp, 2009] Naef, M. and Schupp, J. (2009). Measuring Trust: Experiments and Surveys in Contrast and Combination. *SSRN Electronic Journal*.

[Oliphant, 2006] Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). 'Why Should I Trust You': Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 1135–1144, New York, New York, USA. ACM Press.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei,

L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

[Samek et al., 2017a] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Muller, K.-R. (2017a). Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673.

[Samek et al., 2017b] Samek, W., Wiegand, T., and Müller, K.-R. (2017b). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv preprint arXiv:1708.08296*.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

[Shrikumar et al., 2017] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences.

[Siau and Wang, 2018] Siau, K. and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2):47–53.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

[Turing, 1950] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, LIX(236):433–460.

[Ullman and Malle, 2019] Ullman, D. and Malle, B. F. (2019). Measuring Gains and Losses in Human-Robot Trust: Evidence for Differentiable Components of Trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 618–619. IEEE.

[Verberne et al., 2012] Verberne, F. M. F., Ham, J., and Midden, C. J. H. (2012). Trust in Smart Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(5):799–810.

[Yaochu Jin and Sendhoff, 2008] Yaochu Jin and Sendhoff, B. (2008). Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3):397–415.

[Zeiler and Fergus, 2013] Zeiler, M. D. and Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *European conference on computer vision*.

[Zeiler et al., 2011] Zeiler, M. D., Taylor, G. W., and Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*, pages 2018–2025. IEEE.

[Zhang and Zhu, 2018] Zhang, Q.-s. and Zhu, S.-c. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39.

[Zhou et al., 2015] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Learning Deep Features for Discriminative Localization. *Proceedings of the IEEE conference on computer vision and pattern recognition.*

A

Framework Usage

In order to facilitate the usage of the system, the runner accepts arguments.

- *-m*: defines the mode used by the framework; name of the class containing the run-function
- *-f*: used to pass the name of an image to be processed
- *-all*: can be used instead of *-f* to have all images in the input folder processed

A complete list of all arguments including their description can be obtained using the parameter *-h*.

The input images need to be placed in the folder input. The repository includes two example modes: VGG16 combined with Grad-CAM as well as VGG16 and Grad-CAM++. The missing frozen neural network file vgg16.npy can be downloaded from *https://drive.google.com/drive/folders/0BzS5KZjihEdyUjBHcGFNRnk4bFU* and must be placed in the modes\models folder.

The framework can be run with the command below and arguments where needed.

```
python runner.py –m modeName [-f fileName] [-all]
```

In order to create new modes, it is recommended to model them after the existing *vgg16_gradcampp.py* as well as adding a utils-file. Since the output depends on its purpose, the only requirement for the mode is for the main-class to include a run-function that can be used by the framework to execute the code.

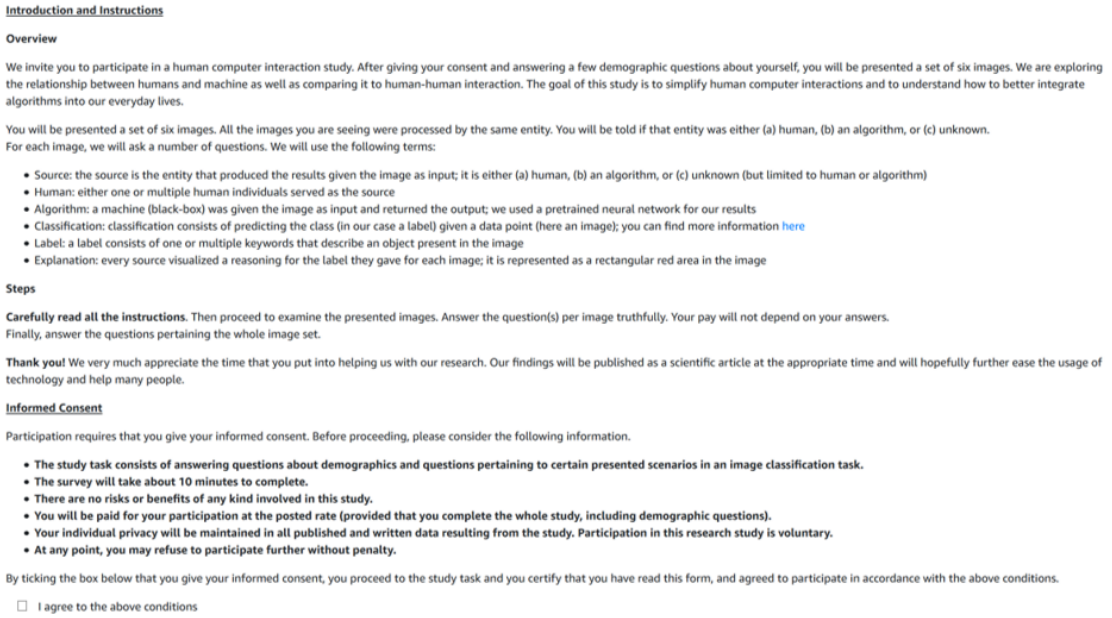
B

The Survey

B.1 Overview

In this appendix, we document the survey designed for the experiment with screenshots.

- Figure B.1: introduction, instructions, and informed consent
- Figure B.2: demographics
- Figure B.3: Affinity for Technology Interaction (ATI) scale [Franke et al., 2019]
- Figure B.4: instructions along with the first image, label, explanation, and questions
- Figure B.5: set of all six images and the first three items of the Multi-Dimensional Measure of Trust (MDMT) [Ullman and Malle, 2019]
- Figure B.6: final three questions regarding trust in strangers [Naef and Schupp, 2009]



Introduction and Instructions

Overview

We invite you to participate in a human computer interaction study. After giving your consent and answering a few demographic questions about yourself, you will be presented a set of six images. We are exploring the relationship between humans and machine as well as comparing it to human-human interaction. The goal of this study is to simplify human computer interactions and to understand how to better integrate algorithms into our everyday lives.

You will be presented a set of six images. All the images you are seeing were processed by the same entity. You will be told if that entity was either (a) human, (b) an algorithm, or (c) unknown.
For each image, we will ask a number of questions. We will use the following terms:

- Source: the source is the entity that produced the results given the image as input; it is either (a) human, (b) an algorithm, or (c) unknown (but limited to human or algorithm)
- Human: either one or multiple human individuals served as the source
- Algorithm: a machine (black-box) was given the image as input and returned the output; we used a pretrained neural network for our results
- Classification: classification consists of predicting the class (in our case a label) given a data point (here an image); you can find more information here
- Label: a label consists of one or multiple keywords that describe an object present in the image
- Explanation: every source visualized a reasoning for the label they gave for each image; it is represented as a rectangular red area in the image

Steps

Carefully read all the instructions. Then proceed to examine the presented images. Answer the question(s) per image truthfully. Your pay will not depend on your answers.
Finally, answer the questions pertaining the whole image set.

Thank you! We very much appreciate the time that you put into helping us with our research. Our findings will be published as a scientific article at the appropriate time and will hopefully further ease the usage of technology and help many people.

Informed Consent

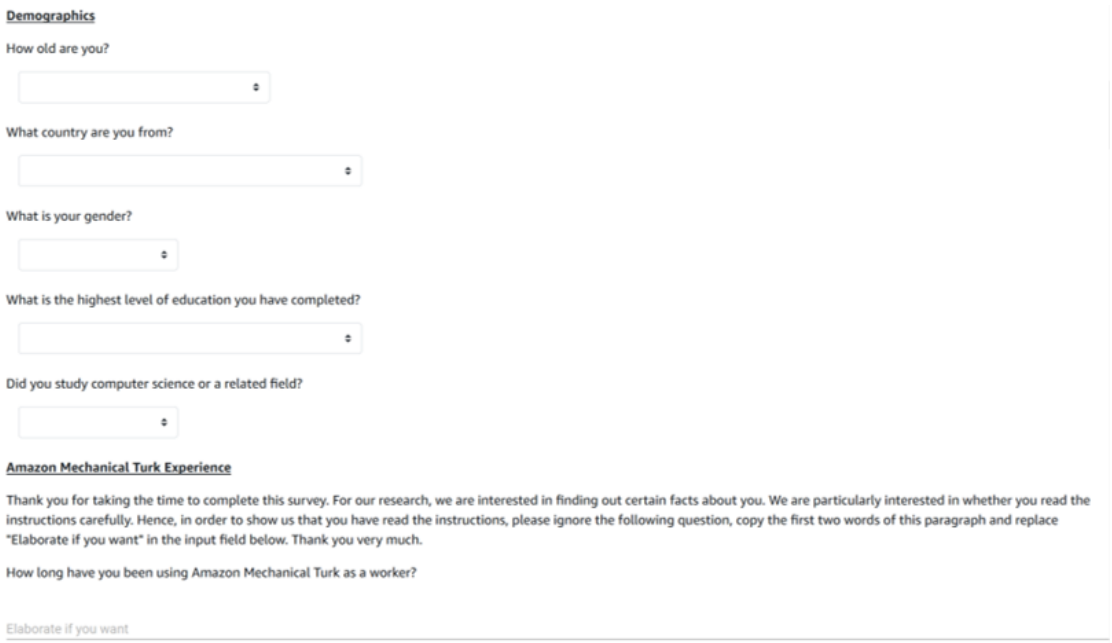
Participation requires that you give your informed consent. Before proceeding, please consider the following information.

- **The study task consists of answering questions about demographics and questions pertaining to certain presented scenarios in an image classification task.**
- **The survey will take about 10 minutes to complete.**
- **There are no risks or benefits of any kind involved in this study.**
- **You will be paid for your participation at the posted rate (provided that you complete the whole study, including demographic questions).**
- **Your individual privacy will be maintained in all published and written data resulting from the study. Participation in this research study is voluntary.**
- **At any point, you may refuse to participate further without penalty.**

By ticking the box below that you give your informed consent, you proceed to the study task and you certify that you have read this form, and agreed to participate in accordance with the above conditions.

☐ I agree to the above conditions

Figure B.1: Introduction, instructions, and informed consent



Demographics

How old are you?

What country are you from?

What is your gender?

What is the highest level of education you have completed?

Did you study computer science or a related field?

Amazon Mechanical Turk Experience

Thank you for taking the time to complete this survey. For our research, we are interested in finding out certain facts about you. We are particularly interested in whether you read the instructions carefully. Hence, in order to show us that you have read the instructions, please ignore the following question, copy the first two words of this paragraph and replace "Elaborate if you want" in the input field below. Thank you very much.

How long have you been using Amazon Mechanical Turk as a worker?

Elaborate if you want

Figure B.2: Demographics

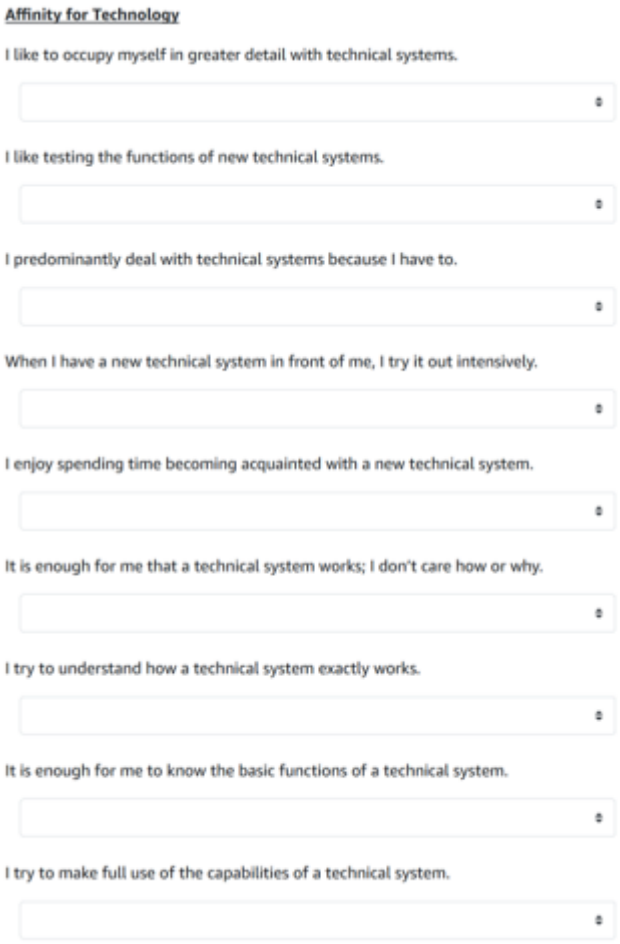


Figure B.3: ATI scale

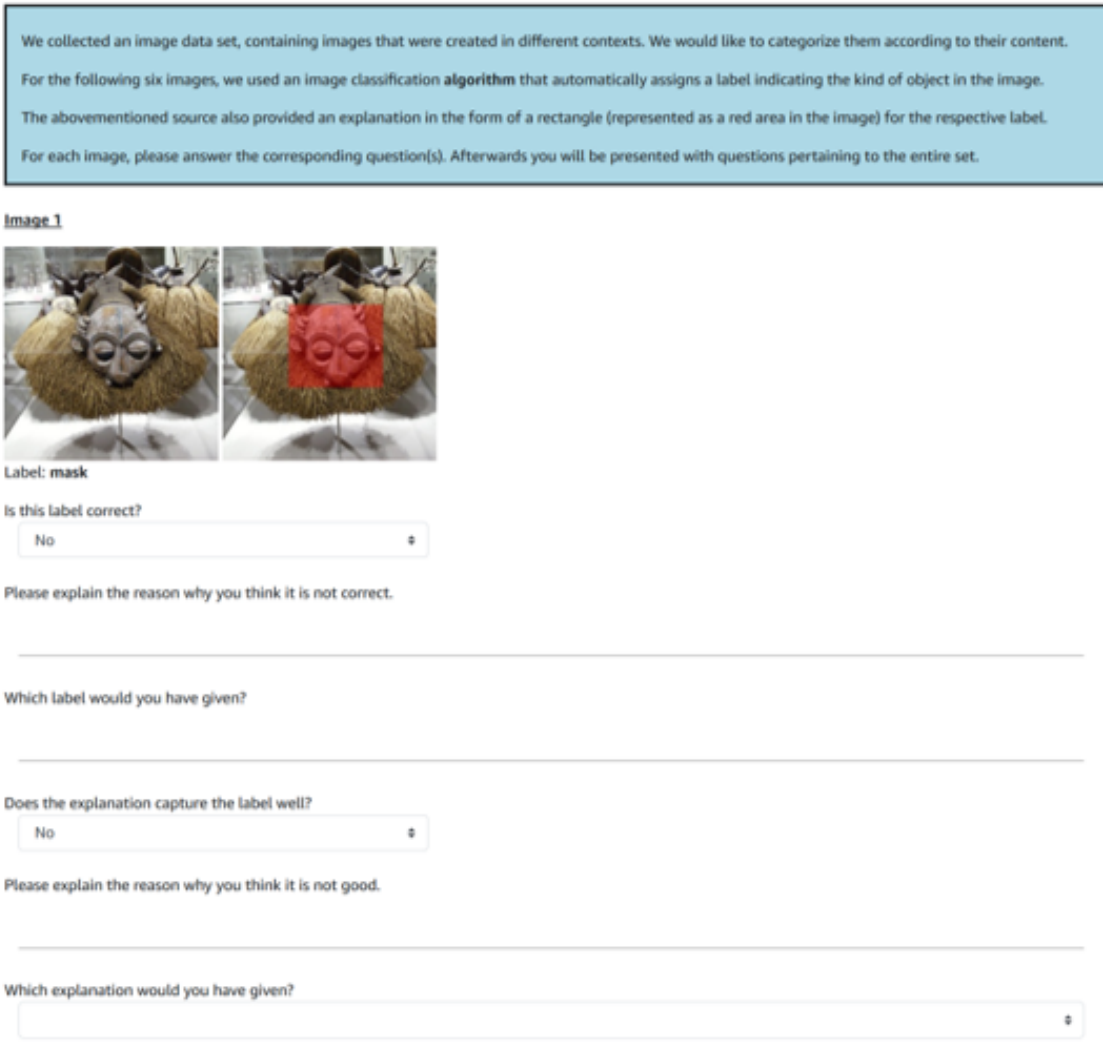


Figure B.4: Instructions along with the first image, label, explanation, and questions

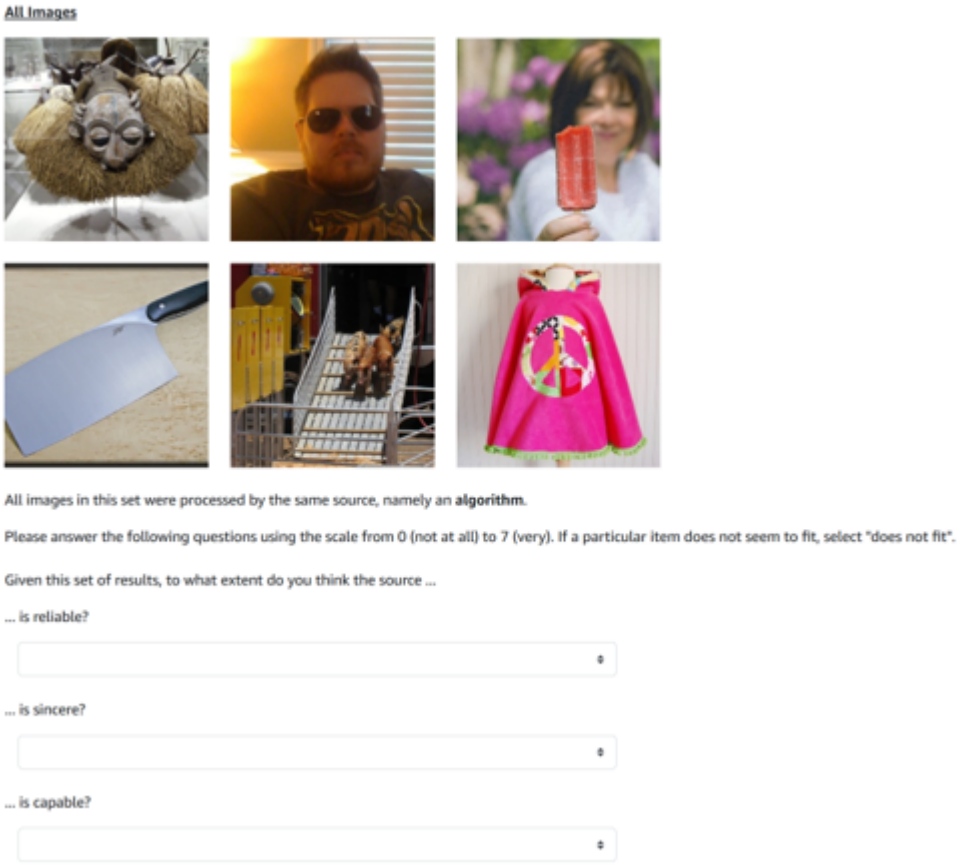


Figure B.5: Set of all six images and the first three items of the MDMT

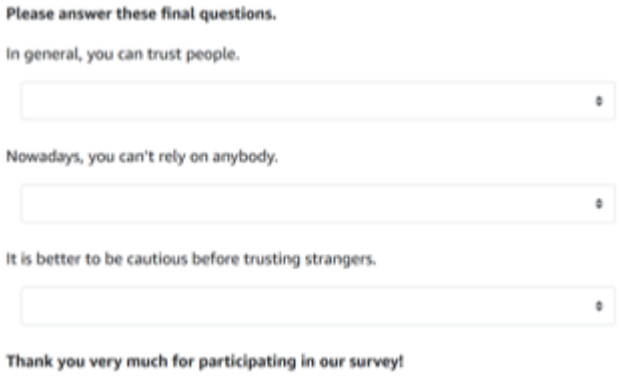


Figure B.6: Final three questions regarding trust in strangers

B.2 Answer Options

In order not to clutter the screenshots, we do not show the expanded drop-down answer options. Rather, we give them down below for questions excluding those that require text as input or are obvious (such as yes/no).

- How old are you?
  - 17 or younger
  - 18 - 20
  - 21 - 29
  - 30 - 39
  - 40 - 49
  - 50 - 59
  - 60 or older
- What is the highest level of education you have completed?
  - Less than high school degree
  - High school degree or equivalent (e.g., GED)
  - Some college but no degree
  - Associate degree
  - Bachelor degree
  - Graduate degree
- ATI Scale
  - Completely disagree
  - Largely disagree
  - Slightly disagree
  - Slightly agree
  - Largely agree
  - Completely agree
- Which explanation would you have given?
  - I would have made the selected field smaller.
  - I would have made the selected field bigger.
  - I would have moved the selected field.
  - I would have moved the selected field and changed its size.

- MDMT: scale from 0 (not at all) to 7 (very) and additionally 'does not fit'
  - 0 (not at all)
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7 (very)
  - does not fit
- Trust in strangers
  - Disagree strongly
  - Disagree somewhat
  - Agree somewhat
  - Agree strongly

C

Testing for Normal Distribution

Table C.1 reports the p-values for normal distribution obtained with SciPy's normaltest[1] of all samples used in the evaluation. All the numbers in the table are rounded to three decimal places.

Dimension / Factor	**[1, 2]**	**[3]**	**[4, 6]**	**[8, 9]**
Reliable	7.236e-02	5.261e-03	2.181e-01	5.336e-01
Capable	7.228e-02	2.238e-03	3.888e-01	3.920e-01
Ethical	1.485e-03	4.648e-01	3.236e-02	3.639e-02
Sincere	5.873e-02	9.150e-04	4.032e-01	1.847e-01
Capacity	7.734e-03	4.840e-04	3.912e-01	7.246e-01
Moral	1.329e-02	9.130e-03	2.466e-01	1.894e-02
Dimension / Factor	**[4]**	**[6]**	**[4, 7]**	**[5, 6]**
Reliable	3.707e-04	3.540e-02	1.572e-03	2.230e-03
Capable	1.837e-04	9.962e-03	8.166e-03	1.164e-05
Ethical	8.056e-04	6.792e-01	3.067e-04	2.555e-02
Sincere	1.224e-05	1.776e-02	2.811e-03	3.403e-04
Capacity	7.217e-05	2.042e-02	3.905e-03	6.048e-05
Moral	2.426e-04	1.091e-01	2.554e-02	1.610e-03
Dimension / Factor	**[5]**	**[8]**	**[7]**	**[9]**
Reliable	7.236e-02	5.261e-03	2.181e-01	5.336e-01
Capable	7.228e-02	2.238e-03	3.888e-01	3.920e-01
Ethical	1.485e-03	4.648e-01	3.236e-02	3.639e-02
Sincere	5.873e-02	9.150e-04	4.032e-01	1.847e-01
Capacity	7.734e-03	4.840e-04	3.912e-01	7.246e-01
Moral	1.329e-02	9.130e-03	2.466e-01	1.894e-02

Table C.1: p-values for the normal distribution of the samples

[1] *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html*

D

Contents of the CD

The CD contains the following files:

- German abstract (zusfsg.txt)
- English abstract (abstract.txt)
- Master's thesis (masterarbeit.pdf)
- Archive of the code repository (2019-florian-ruosch.zip)

List of Figures

List of Tables