UI / UX requirements for explainable and trustful AI systems

1st Project Thesis

submitted on August 27, 2020

Faculty for Economics

Business Informatics for Data Science

WWI2019F

by

Tobias Hölzer

Supervisor at the Company:

IBM
Svenja Laing
Software Developer

DHBW Stuttgart:

Dr. Kai Holzweißig

Signature of the Supervisor:

The rising field of Artificial Intelligence (AI) is developing very fast. The use of AI comes with a lot of risks since the reasoning processed of an AI are very hard to understand which concludes into overlooked or unfounded mistakes made by the AI. Unfounded mistakes leads often to reduced trust into an AI. Therefore it is important to help the user to understand AI and increase the trust into such systems. Visualization approaches seem to be a helpful factor when understanding complex topics like AI.

This work combines existing principles for developing a trustful AI from different institutions and companies into six trustful principles. It declares some requirements for creating a visualization which increases trust into an AI system based on the combined trustful principles. These requirements are then used to develop a concept and a prototype for an example AI system.

Contents

List of abbreviations

IBM International Business Machines

IEEE Institute of Electrical and Electronics Engineers

AI Artificial Intelligence

QPSO Quantum-based Particle Swarm Optimization

WDPS Water Damage Prevention System

UI User Interface

UX User Experience

IT Information Technology

ML Machine Learning

Carbon IBM Carbon Design Component System

List of Figures

List of Tables

1 Introduction and motivation

In the past years the subject area around AI developed very quickly. The usage of AI comes with a lot of risks. Some use cases of an AI system are very critical to false predictions, like medicine. For example, if an AI should predict the disease of the patient. If the AI predicts the wrong disease and the responsible doctor believes the prediction of the AI, the patient could get the wrong medicines or treatment. So, the doctor needs to understand the AI and its decision to choose for each case weather he believes the AI or not.[1] Another example is the prediction of potential criminals in the United States. A police station trained an AI to predict whether an individual could be criminal or not. For the training they used lots of personal information, between others the skin colour. Because the United States has a huge problem with race inequality, people of colour get more often arrested than the average american and the AI thought that people of colour are more criminal than average american. Lot of statistics shows that that is a false conclusion from the data.[2] AI systems are not always bias free, because the data behind them are not bias free. Therefore, it is very important that the people which develop, work and are influenced by AI systems have access to the topic and a healthy relationship to the AI.

Most AI systems can be described as black-box systems, i.e. the reasoning procedures AI systems use to make decisions are mostly hidden or too complex to understand. Without being able to look inside an AI system and understand why an AI system came up with a conclusion, many people do not trust nor use AI based solutions. To give people access to this increasingly complex topic, many companies and institutions, like Institute of Electrical and Electronics Engineers (IEEE), International Business Machines (IBM) and Google, developed programs and conducted scientific research about how to improve the relationship between humans and AI. The main focus of this research is how to make AI systems more trustful.[3][4][5][6][7]

1.1 Storyline of the Water Damage Prevention System as a practical use case

Buildings can take damage worth many millions of euro through pipe breaks and leaks. Therefore, the developer team at the Watson Center attached humidity sensors all over the 20th and 21st floors to be aware of such breaks in time. A dashboard will raise alarm if one sensor measures a humidity-value over a specific threshold. This method is not very accurate since the measured

[1] Cf. Ribeiro/Singh/Guestrin 2016, p. 1136
[2] Cf. Will Douglas Heaven 2017, para. 7 et seq.
[3] Rossi et al. 2019
[4] IEEE 2017
[5] Google AI 2018
[6] Microsoft AI 2020
[7] EY 2018

humidity-values changes with many factors like the weather or also indoor activities like mopping the floor. On rainy days the overall humidity raises also indoors and that may cause an alarm if the threshold was set to low. On the other hand, if the threshold is set to high it may occur that on a very dry day no sensor humidity-value will reach the threshold and a breach will remain undiscovered.

To encounter this problem the team implemented a Water Damage Prevention System (WDPS) which uses a Quantum-based Particle Swarm Optimization (QPSO) algorithm to dynamic identify a possible leak. The WDPS is an application which monitors the humidity values in buildings with an array of sensor to detect water leakages and sends notifications if a water leakage is detected. To detect water leakages the swarm intelligence searches for anomalies in a virtual landscape generated by the values of the sensor array. This allows the user of the WDPS to prevent water damage more accurate without high effort and costs.

Yet, there is no visualization of the WDPS which would help the user to understand the very complex system of the QPSO algorithm. Potential user would not be able to understand the system and its results and predictions. That makes the system untrusted and therefore not useful for humans. A visualization which explains the system would increase trust in it. The visualisation would increase the relationship between humans and the WDPS and explicit the AI system behind it. To support the visualisation an User Interface (UI) with a good User Experience (UX) is required. *But what aspects of UI and UX are necessary to make a system with AI explainable and trustful?*

To answer this question, I evaluate and analyse other examples of trustful and explainable AI systems, collect findings of other research, develop a visualisation for the WDPS under the best aspects I found and analyse it as an artefact for this study. With the analysis and evaluation of the artefact it is possible to either declare that the found aspects are necessary aspects of UI and UX to make a system with AI explainable and trustful or to conclude that the found aspects are not necessary to do so.

2 Methods

Whether an user experience an UI and UX well or not is strongly connected to personal preferences and opinions about design. That is the reason why it is hard to use meaningful quantitative research methods to verify design hypotheses and why this thesis will use mostly qualitative methods. The frame of this research will be the design science approach. To find relevant criteria and examples in literature this thesis will conduct a structured literature review. A concept and a prototype for a 3D visualization of the swarm intelligence behind the WDPS based on this concept will be designed and developed to evaluate the found relevant criteria. This prototype should then be the focus of of a survey about explainability of AI systems with the example of the WDPS to validate whether the found criteria are helpful to design trustful AI or not.

The goal of scientists and researchers which approach a design science research is to have developed an artifact at the end of the research. This artifact is developed in several research iterations in which the researcher evaluates the recent iteration artifact and develop a new one from the new knowledge.[8] To apply that approach to this research it is important to define critical knowledge in the topic of AI first, with which it will be possible to focus the review literature on important points about trustful AI. Some hypotheses will formulated in the literate review, which will be the scientific base of the resulting artifact; a concept and a prototype of this concept. The developing process of the concept is iterativ. Firstly the concept is evaluated based on the found criteria for a good visualization. Secondly the concept is refined based on the results of the first evaluation. Based on the second concept a prototype will be developed which will help in cooperation to proof the hypotheses of the literature review.

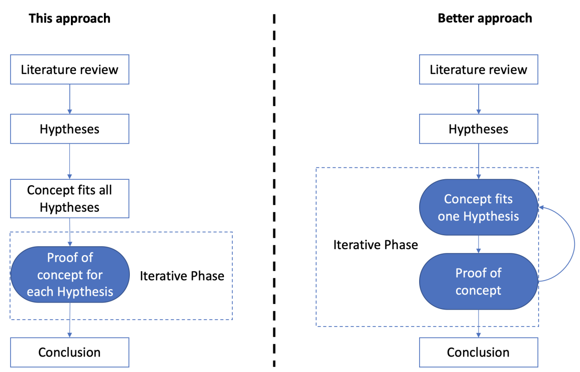

Figure 1: Left side: concept of own approach with only one Design Science iteration for all hypotheses. Right side: concept of potential better approach with Design Science iterations for each hypothesis. [9]

The described research methods are not the best for doing research about this topic. Due to time constraints only parts of the Design Science research approach are applied with some limitations. A better approach would be to iterate through each hypotheses and develop an artefact to each

[8] Cf. Holmström/Ketokivi/Hameri 2009, p. 67 et seqq.
[9] Design Science

of them. [10] Another problem are missing capacities to do a user survey to validate the prototype. Therefore it is not possible to have a scientifically prooven conclusion.

[10]See Figure 1

3 Theoretical background of Artificial Intelligence

AI is a commonly used term, but its definition is not set correctly. Most people do not even know how far technology reached so far in developing AI. In movies AI is often seen as Robots becoming the rulers of the world and supress the human race. It seems like humans distrusting AI and AI driven systems by nature, even if it is not very clear what exactly can be defined as an AI. What makes a program, e.g. the QPSO algorithm, intelligent and what is meant by trustful AI?

3.1 Artificial Intelligence

The definition of AI changed a lot in the short Information Technology (IT) history. The name "Artificial Intelligence" appeared the first time in 1956 on the Dartmouth Summer Research Project on Artificial Intelligence conference. Before that conference AI was known as computer intelligence or machine intelligence. The first paper ever written about the topic was a typewritten paper from Alan Turing in 1941.[11] But since then the field moved on and redefined itself over and over again because the term intelligence itself is hard to define. One definition of intelligence was made by Lloyd Humphreys: "[intelligence is] the entire repertoire of acquired skills, knowledge, learning sets, and generalized tendencies considered intellectual in nature that are available at any one time".[12] Todays AI systems are not corresponding to this definition, because they have very specific use-cases. Natural Language Processing for example is marked as intelligent but such systems cannot debate about politics. Their kind of intelligence is not useful in every situation. It seems like there are different definitions for each development state of AI systems.

To handle with every development state of AI, the definitions of it can be classified into four categories:[13]

- Systems that think like humans
- Systems that act like humans
- Systems that think rationally
- Systems that act rationally

A definition of AI was made by Dimiter Dobrev: "**AI will be such a program which in an arbitrary world will cope not worse than a human.**"[14] That fits with the definition

[11]Cf. Moor 2003, ch. 1 p. 1 et seqq.
[12]Humphreys (1971) quoted by Scarr 1989, p. 74
[13]Cf. W Boers et al. 2009, p. 2
[14]Dobrev 2004, p. 2

of the first development state of AI systems: AI systems are amongst other "systems that act rationally".[15]

3.2 Swarm Intelligence

Swarm intelligence describes the field of biotic problem-solving methods which are based on or inspired by natural swarm structures. The concept behind swarm intelligence is based on single elements which are acting chaotic and random alone but acting structured when moving together. Each unit of a swarm follows specific rules and the units join together to a swarm. This swarm then acts like it thinks rational and can be used to solve specific problems. Swarm intelligence are sometimes used as optimization algorithms for example the QPSO algorithm.[16]

A good example for a swarm intelligence is a bee swarm. Sometimes a bee swarm must search for a new place to build their hive. This place must be well chosen, there are many requirements for the optimal place. One bee alone cannot find that place alone, but all bees together can. They spread out and searching for the place, if a bee thinks that it has found a nice spot it tells the others. Then the other compare the place with their own best place found so far and if tells others which place is better. That way the swarm has better chances in finding a good spot for their hive that a single individuum has.[17]

Another example is the decision making of a human group. A group of five people want to order some food. Of course, they can vote weather they want to order pizza, sushi, Mexican or Indian food. Two people votes for sushi, each other possible option has exactly one vote. Two of the other three people who did not voted for sushi does not eat fish because they are vegetarian, the other just do not like fish. But everybody likes pizza, not a favourite food but all like it. In that example the option to order sushi wins the vote, but it is not the best option for the group, which would be to order pizza. Scientist developed a system which helps finding in such cases the right decision. With that system the human group can act like a swarm, each individual has a best position and shares it with the others. Some best positions are speaking against each other, e.g. the vegetarian position speaks against the sushi position. Each individual can now react to the others best position and change it if necessary. If three people position themselves against sushi, the other two are outvoted and start to match their position to another food what they like, e.g. pizza. That way the group can find the optimal decision with the help of a swarm approach.[18]

Referring to the definitions of an AI a swarm intelligence like the QPSO would be artificial intelligent, because its actions are based on rational created information from environmental data and the execution of these actions would not respond to having more problems reacting to its environment than human actions would.

[15] W Boers et al. 2009, p. 2
[16] Cf. Kennedy 2006, ch. 6 p. 1 et seqq.
[17] Cf. Karaboga/Akay 2009, p. 64 et seq.
[18] Metcalf/Askay/Rosenberg 2019

3.3 Trustful Artificial Intelligence

As already mentioned, humans have a natural distrust against AI systems. This distrust is represented through dystopian movies and books like Matrix. This kind of distrust in addition to the rapid development of AI in the last decades resolved in a societal discussion about how to handle with AI systems.[19] Also, todays AI systems are very vulnerable to biases which leads to false recommendations and predictions. That results in even less acceptance and reliability on AI systems by humans. To encounter this problem many companies and institutions doing research on how to make AI systems more reliable.

Because there are so many people doing research on the topic many different terms for reliable AI systems occurred, e.g. trustful AI from IBM or responsible AI from Google. They all describe the same concept of a more reliable, responsible and trustworthy AI to increase the acceptance on AI systems by humans. I will continue using the term “trustful AI” for this concept of a reliable / responsible / accepted AI.

The companies and institutions IBM, Google, IEEE, Microsoft and EY defined some principles and practices for creating trustful AI.[20][21][22][23][24] Compared to each other, the most important principles seem to be: [25]

- Fairness
- Explainability
- Transparency & Accountability
- Security & Privacy
- Social Responsibility
- Technical Availability

The most important point of the fairness principle is that AI systems should be bias-free. Biases are very dangerous in science, but also in public since they are difficult to spot, may result in wrong research results and potentially decrease the life quality of individuals if their social environment is toxic through biases.[26] According to IBM Research there are more than 180 human biases identified and classified.[27] According to some fairness principles, it is important to include everybody as a potential stakeholder. A good example for that could be an intelligent dating algorithm which connects strangers together and arrange a date for them. An usual

[19] E.g. Purdy/Daugherty 2016
[20] Rossi et al. 2019
[21] IEEE 2017
[22] Google AI 2018
[23] Microsoft AI 2020
[24] EY 2018
[25] See Table 1
[26] E.g. Nguyen-Kim 2020, at 2:55 seq.
[27] C.f. IBM Research 2018, at 0:06

	Fairness	Explainability	Transparency & Accountability	Security & Privacy	Social Responsibility	Technical Availability
Microsoft						
Fairness	X					
Reliability & Safety				X		X
Privacy & Security				X		
Inclusiveness	X				X	
Transparency			X			
Accountability			X			
Google Principles						
Be social beneficial					X	
Avoid creating or reinforcing unfair bias	X					
Be build and tested for safety				X		
Be accountable to people			X			
Incorporate privacy principles				X		
Uphold high standards of scientific excellence					X	
Be made available for uses that accord with these principles					X	X
Google Practices						
Fairness	X					
Interpretability		X				
Privacy				X		
Security				X		
IEEE						
Human Rights	X				X	
Well-being					X	
Data Agency			X			
Effectiveness						X
Transparency			X			
Accountability			X			
Awareness of Misuse	X					
Competence						X
EY						
Performance						X
Resiliency				X		
Bias	X				X	
Explainability		X				
Transparency			X			
IBM						
Transparency & Accountability			X			
Value Alignment					X	
Explainability		X				
Fairness	X					
User Data Rights				X		
Robustness				X		
SUM	8	3	8	9	8	5

Table 1: Comparison of trustful AI principles of different companies. Each company principle for trustful AI can be mapped to at least one shared principle.

practice would be starting with including the connection of hetero males with hetero females in the algorithm and after that include other sexual groups into the algorithm. But that could result with forgotten minorities or even excluded minorities. A better approach would be tinking about "How we can design for the 3%. We can then solve fot the 97% at the same time."[28] That way it is ensured that everybody is included as a potential stakeholder of the AI system. The fairness principle also describes that AI systems should consider human rights in their development and execution as well as they should be aware of potential misuse and try to prevent the system from that.

The approach of the explainability, the transparency and accountability principles is to increase the users understanding of the AI and its reasoning process, because humans often distrust what they do not understand. The principle explainability describes that an AI system should explains itself to an user so good, that an user can interpret these explanations easily. So, explainability comes only with interpretability and with this interpretability. But for a solid understanding of the AI system it is also important that the user can have a transparent look at the data the AI is consuming. With this information he can ensure that no obvious error during the reasoning process occurred caused by faulty data and check if the output of the AI correlates with the consumed data. Also, the AI system must be able to account on the data it consumed, so the user can trust into the correctness of the consumed data.

Visualizations are perfect for applying the explainability, the transparency and accountability principles. Humans often understand certain topics better if they are explained with the help of visualizations. The subject of Data Visualization examines how exactly visualizations can increase the understanding of certain topics. With Visualizations of the consumed data it is also possible to get more transparency into an AI system. Only the accountability cannot be improved by visualizations, there should be other technologies for that principle.

Security and privacy is also included as a fundamental principle in designing AI systems. Because AI systems could be used in critical environments they need to be secured from potential cyber-attacks like poisoning or evasion.[29] AI systems consume often personal data; therefore, it is important to ensure privacy rights and secure these.

AI systems also should be socially responsible and have a broad technical availability. According to Google and IEEE an AI must always be social beneficial and improve social well-being.[30][31] They also should be performant and effective like software always should be. But for AI systems it is also very important that they work with today's standard technologies and not with scientific or future ones. AI systems should exist for the people.

[28] Microsoft AI 2020, in 'Inclusiveness' at 0:29
[29] E.g. Polyakov 2019
[30] Google AI 2018
[31] IEEE 2017

4 Analysis of visualizations approaches in literature

To develop a concept for a UI/UX prototype it is helpful to find some potential requirements and examples in the literature and define some new requirements based on edge cases for this use case. In the following are examples in literature and free web of AI, sensor and Machine Learning (ML) visualizations described and analysed to extract requirements to a UI/UX.

4.1 Example based explanations with Quickdraw

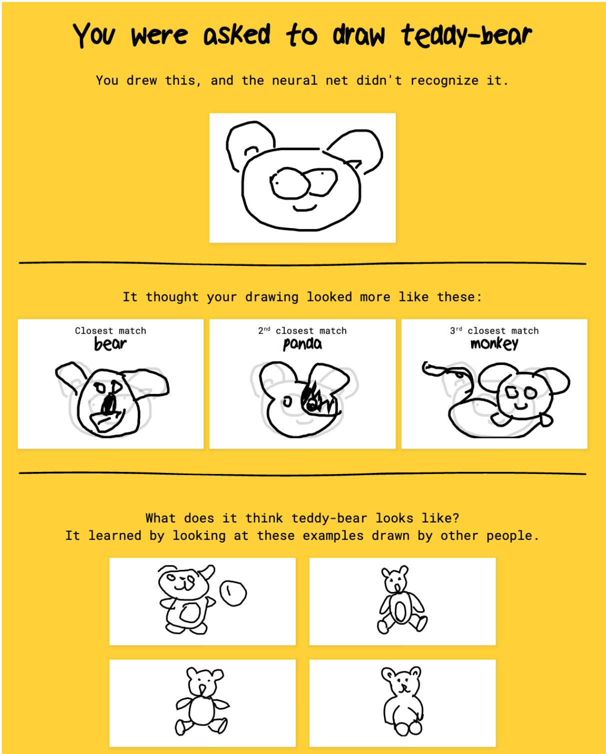

Figure 2: Quickdraws explanation of its results from the drawing to the word teddy-bear. The explanation is splitted into three sections from top to bottom. The drawn picture from the user is shown at the top. Under the drawn picture is the comparative explanation which shows the user which other pictures of words looked similar to his drawing. At the bottom is the normative explanation visible. The normative explanation shows drawings of the same word from other users.[32]

Quickdraw is a game like AI experiment from Google in which the user should draw a given word. At the beginning of each round the computer says a word which the user should then draw on the screen. The user has 20 seconds time for that. While the user is drawing the word, an AI tries to guess that word. If the AI guess the word right or the 20 seconds passed the user should draw another word. After six words the game ends and the UI show a retrospective of the game.

[32] Quickdraw Explanations

The user can get an explanation for the results of the AI for each word.[33] The explanation for the AI results are spliced into two categories: a comparative explanation at the top and a normative explanation at the bottom. The normative explanation shows drawings from other people which where categorized correctly from the AI. That way the user can compare his own drawing with other ones and maybe see why the AI did or did not guessed the right word. The comparative explanation shows other words drawings which are similar to the drawing from user. With the comparative explanation the user is able to understand why the AI guessed other words in the first place. Referring to this example a study from 2019 shows that normative explanations increase the users rating of capability & understanding of an AI system. Comparative explanations decrease the users rating of capability of an AI systems.[34] However, the team behind that study mentioned in their discussion that this phenomenon could have a positive effect on the trust given to AI systems: "[. . .] different explanations could be leveraged depending on the goal. For instance, if users are known to under-trust a system despite high system capability, normative explanations could help improve user perceptions during system errors. Alternatively, if users tend to over-trust the system, comparative explanations could help establish a more appropriate level of trust."[35] According to their study, the use of normative and comparative explanations of AI systems could help generating a healthier relationship between humans and AI.[36]

Hypothesis: *Normative and comparative explanation UI/UX elements are important for the relationship between humans and AI. They also increase the explainability of an AI system and therefore the users trust in it.*

4.2 Explaining a prediction ("Why should I trust you?")

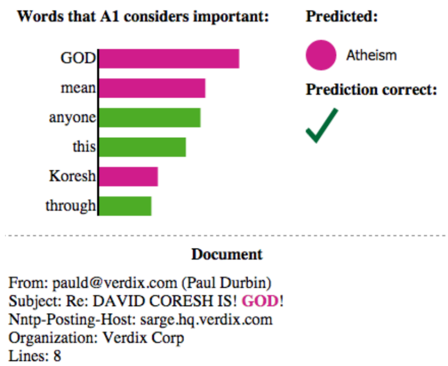

Figure 3: Visualization approach to explain the predictions of a specific classifier whether a document is about "Christianity" or "Atheism". The relevance of a word is measured by the bars lengths, the assigned class is represented by the bars colour. Image from Ribeiro/Singh/Guestrin 2016, p.2.[37]

[33] See Figure 2
[34] Ct. Cai/Jongejan/Holbrook 2019, p. 3 et seq.
[35] Ct. Cai/Jongejan/Holbrook 2019, p. 4 col. 2
[36] Ct. Cai/Jongejan/Holbrook 2019, p. 4 et seq.
[37] Decision Weights

Another explanation approach is the weighted relevance visualization. A team of the University of Washington tried to automatic explain any classifier with their new LIME technique.[38] Classifiers are AIs which classify data into specific groups on the basis of e.g. decision trees. Their new technique gives every data attribute, which led to a prediction, a relevance weight. This weight is either negative, means this attribute value speaks against the prediction, or positive, means this attribute value speaks for the prediction. The amount of the relevance weight describes how strong a data attribute speaks for or against a prediction. Their technique also filters irrelevant data attributes out to decrease the amount of potential biases. If an AI should predict which disease a patient has, some data attributes like the age are irrelevant for the prediction. Sometimes the age is important, for example if only children can get a specific disease. But for other diseases like the flu the age does not matter. Of course, older and very young people get more often infected by the flu, but the age is not an indicator for it. So, remove irrelevant data attributes can decrease the amount of potential errors. Also, the user would not get irritated by irrelevant information if these data attributes gets removes.[39]

Hypothesis: *A key point in explaining predictions of AI systems is the quantification of the decision weights of data attributes. That way it is possible for the AI system and for the user to work with concrete numbers and therefore a better possibility of comparing different states with each other.*

4.3 Sensor Network Visualization

The environment of the swarm is generated based on the data of a sensor network. The swarm with its particles can also be seen as a network, at least the resulting data could follow visualization rules of network data. So, all this network data must also be visualized. According to B. Shneiderman visualizing network data comes in seven tasks:[40]

- Overview: Gain an overview of the entire collection.
- Zoom: Zoom in on items of interest.
- Filter: Filter out uninteresting items.
- Details-on-demand: Select an item or group and get details when needed.
- Relate: View relationships among items.
- History: Keep a history of actions to support undo, replay, and progressive refinement.
- Extracts: Allow extraction of sub-collections and of the query parameters.

[38] Cf. Ribeiro/Singh/Guestrin 2016, p. 1
[39] Cf. Ribeiro/Singh/Guestrin 2016, p. 1 et seqq.
[40] Cf. B. Shneiderman (1996) quoted by Zankl 2009, p. 4 col. 1

The user should be able to gain himself an overview over the data set to gain and understand overall context information about the current situation which is represented by the data. After the user gained himself this context information, he must be able to zoom into specific areas of interest, for example a peak in one's sensors history. Also, the user must be able to get details on demand, to understand specific items of interest. To recognize interesting or important areas and items easier, the application must provide a filter function with which the user can filter out uninteresting items. It should also possible to follow the relationships among different items to get specific context of a data point. As a standard user experience component an activity history should not be missing as well.

Another factor at visualizing sensor network data should be considered is the dimension of the data. A two-dimensional data field can be perfectly shown in a three-dimensional visualization. The positions of the two-dimensional data field filling the x and y parameters, the values of the data points are represented in the z parameter. Other dimensional data fields behave similar: for example, a three-dimensional data field. The positions of a three-dimensional are represented in the x_1, x_2 and x_3 parameters, the value of the data points in the x_4. To represent an x-dimensional data field it needs an $(x+1)$-dimensional visualization.[41]

Hypothesis: *The optimal dimension for a visualization of a network data field is x+1, x means the dimension of the network data field. Therefore, the optimal dimension for a visualization of a two-dimensional network data field is three-dimensional. Including the seven tasks of visualizing network data will improve the user experience. That will also increase the transparency of the used data, because the user is able to gain insights into the data. Including the seven tasks would because of the increase of transparency also increase the trust of the user into the data from the sensor network and therefore the trust in the AI which uses the data. Generating trust in data means generating trust in AI.*

4.4 Intro to Machine Learning by R2D3

R2D3 is a Website from Tony Chu and Stephanie Yee where they give a visual introduction into the topic of ML within two parts. They are using the example of a decision tree to explain basics like the difference between bias and variance. [43] The website uses animations triggered on scroll to explain these topics. Because these topics are not related to a swarm intelligence since swarms do not use ML most parts of this visualization are not transferable on this topic. However, at one specific point they describe how a single data point was categorized into the wrong category. With the help of that single data point and its decision path they were able to find and explain a potential error in their decision tree.[44]

[41] Cf. B. Shneiderman (1996) quoted by Zankl 2009, p. 3 col. 1 et seq.
[42] R2D3 Decision Path
[43] See Figure 4
[44] See Chu/Yee 2015

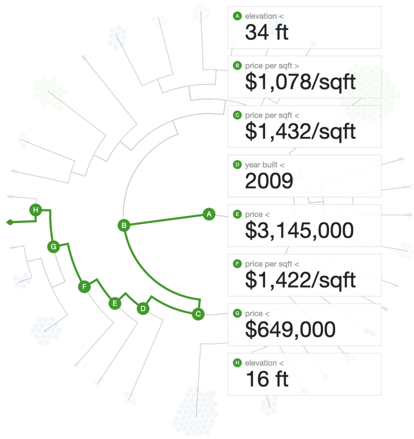

Figure 4: A visualization of a decision path of a decision tree ML algorithm. The algorithm should predict whether a home is located in San Francisco or New York. The visible leaf shows a prediction for a single San Francisco home. With the help of this visualization it was possible for the authors to explain the variance problematic of a ML algorithm.[42]

Hypothesis: *Giving the user the chance to follow the decision path of a single data point increases the explainability of an AI system and therefore the trust into it. Explanations of the AI system by following the decision path of a single data point would help the user to understand the reasoning process of the AI and would therefore it would increase trust into it. It would also help the user to find potential errors within the reasoning process, which would increase the accountability of the system and therefore again the trust in it.*

4.5 Necessary parts of trustful visualization

Requirement increases	**Trustful Factors**
Normative & Comparative Explanations	Explainability
Quantified Decision Weights	Explainability & Accountability
Irrelevance Filter	Explainability
Seven Tasks of visualizing Network Data	Transparency
Precise Data Insights	Explainability & Accountability

Table 2: Collected requirements with their fullfilling principles for a UI or visualization which increases the trust of a user in an AI system.

In the literature it seems that some necessary practices are defined, to create a trustful sensor network-based AI visualisation with a good UI and UX. That requirements could be used to develop a concept and a prototype of a visualization for a sensor-based swarm intelligence AI system.[45]

[45] See Table 2

Hypothesis: *The visualization of the AI system should have normative and comparative parts, which could improve the relationship between humans and AI and also improve the explainability of an AI system. Quantified decision weights with a filter for irrelevant data can improve the explainability and accountability of an AI system. To explain the data consumed and created by an AI system which would increase the user experience and the transparency of an AI system, the visualization should include the seven tasks Overview, Zoom, Filter, Details-on-Demand, Relate, History and Extracts. The visualization of an AI system should also provide precise data insights, e.g. a decision path, because it would increase the explainability, transparency and accountability of an AI system.*

5 Concept and prototype of the visualization for the Water Damage Prevention System

To proof the hypothesis of required practices and techniques for creating an UI/UX for a trustful AI system I created a prototype UI/UX for a sensor-network-based swarm intelligence.

5.1 About the used swarm intelligence

The WDPS uses a swarm intelligence to recognise potential water leaks. This swarm intelligence is a quantum-based particle swarm optimisation (QPSO) algorithm which operates on a virtual surface generated by the humidity sensor values on a floor of a building. Each point of the surface represents a point existing in the real-life building. The virtual surface dents to the bottom if a sensor measures higher values. Therefore, the global minimum of the surface represents the highest humidity on the floor. The swarm searches for the global minimum, which is the best spot for a particle, on the virtual surface. The particles communicate the best spots with others, every particle has an own best spot found in the past and a social best spot, the best spot the swarm has found yet. Each particle is attracted to both points, in addition to its current vector speed. Also, the particles have quantum properties and therefore they are doing sometimes little quantum jumps and acting with a specific randomness factor. When particles contract to a point this point could be the global minimum. This global minimum refers to a potential leak.

5.2 First concept

The UI concept represents a Dashboard which is spliced into four parts: the focused 3D visualization of the sensor network and the swarm intelligence, a filter-field, a meta-data-field for additional information and an entity-field where the user can request detailed information on demand. [47] The 3D visualization has the primary focus on the Dashboard, given by its size. The visualization shows the environmental surface of the swarm. On top of the surface the particles of the swarm are represented by dots. A little bit above zero level is each sensor represented by a dot which size represents its measured value. In this 3D visualization are therefore three entities represented: The swarm with its particles, the swarm environment surface and the real-life sensors. Each entity can be selected by the user.

When an entity is selected the entity-field shows more details about the selected entity. For the swarm the entity-field shows some details about the current contraction of the particle and a history chart of the its confidence level for its decision. The entity-field shows for the surface

[46] First Concept
[47] See Figure 5

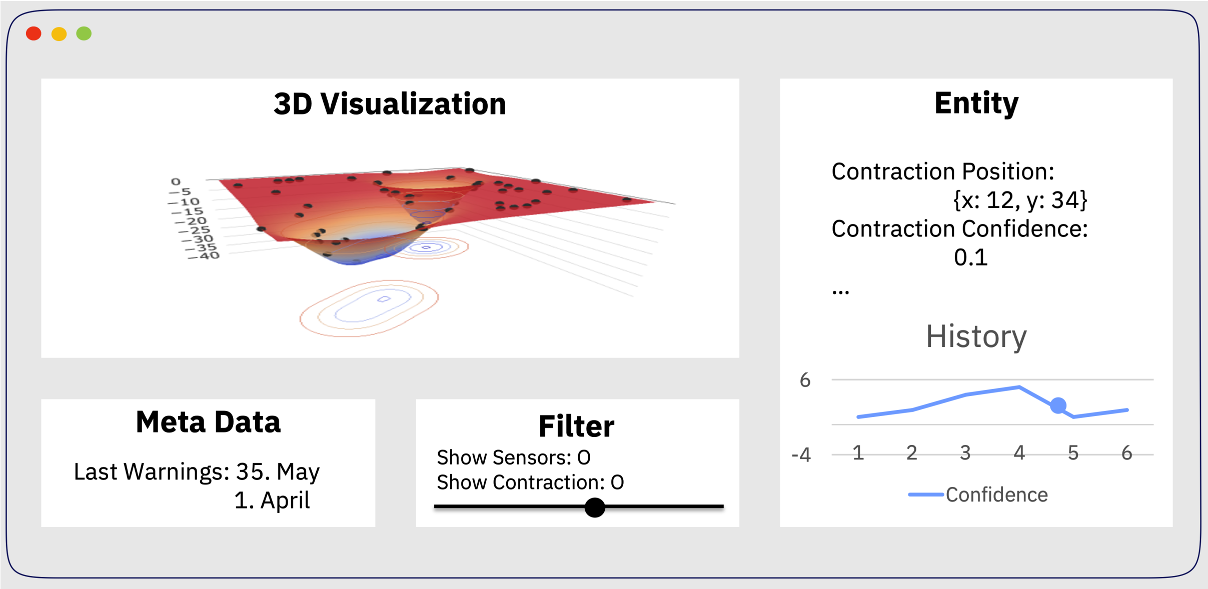

Figure 5: First Concept of the UI for the WDPS. The UI is spliced into four boxes: the large 3D visualization, an entity section, a filter section and a meta section. Each section is marked with its name as box headline. [46]

information about specific points, which position and height they have but also which sensors influence the specific point the most. A history of the recent height is shown in a graph below. Pretty similar data fields can be found for a specific sensor.

At the filter-field the user can filter the shown data, e.g. show and hide entities. But most important the user can select via a history slider a specific time step. Then the 3D Visualization changes to represent the swarm and sensors at the specific time point. With this history slider it is possible to compare different states of the swarm to get information about potential errors.

The meta-data-field gives the user more information about the system additional to some explanations. Information about the system could be recent incidents and alerts and also the current status of the system, whether it thinks that a leak occurred or not.

5.3 Analysis of the first concept in consideration of the trustful principles

The concept of the prototype needs to fulfil the trustful acception criteria: [48]

- Normative & Comparative Explanations
- Quantified Decision Weights
- Irrelevance Filter

[48] See Table 2

- Seven Tasks of visualizing Network Data
- Precise Data Insights

The user can compare the current state of the swarm with older states to be able to see differences between the states. When a warning occurs, the user has the ability to jump to a state where an older warning occurred and compare these two states. That will give the user a normative and a comparative insight about possible errors in the decision making of the AI system. A side-by-side comparison is not included into the concept but that could increase the UX of the use of the normative and comparative explanation. Also, the user must be trained to recognize potential errors in the system, e.g. a wrong or random contraction of the swarm. So, the concept provides a not very normative and comparative explanation of the AI system.

Through the contraction weight of the swarm the AI system provides a quantified decision weight. This contraction weight is calculated at the contraction determination and tells how strong the particles of the swarm are contracted to each other. The more the particles are contracted to each other the higher is the possibility of a global minimum and therefore a water leak. So, this contraction weight can also be seen as a confidence number of the AI system. This confidence number is the only direct quantifier for a decision whether the AI raises an alert or not. Other indirect quantifiers are also decision weights, but they only lead to the confidence number. These quantified decision weights are provided in the entity field. Each surface entity shows the distance to the most influencing sensors, the swarm entity shows a specific contraction point and the distance to the most influencing points on the surface.

The concept provides filter possibilities in the filter field. There the user can filter out the irrelevant data, such as non-contracting particles or sensors which are far distantly from the swarm's contraction. The concept also provides meta information and context about the swarm like recent incidents or alerts, a floor plan or a specific bell-gamma value with which the surface is generated with. Providing such specific values reduce the self explainability of the system but improves the interpretability because an informed user can check if this value agrees with the results of the swarm.

The seven tasks of data visualization are also represented in the concept. The first task overview is given through the 3D visualization. After page load the user can see the complete scene of the swarm, including information about a contraction of the particles and therefore a potential leak. Then the user can zoom into the scene and move around. As already mentioned, the user has the ability to filter out irrelevant data and entities. The most crucial task to provide details on demand is also fulfilled by the concept. On click on an entity the entity field shows more information about the entity and provides details. The two tasks history and extracts are not fulfilled by the concepts. The history task could be added by two buttons for 'undo' und 'redo', the extracts task is at least a little bit fulfilled by the history slider, with which the user can select a specific time step.

The concept does not fulfil all required criteria of a visualization for a trustful AI system. The missing parts are a side-by-side comparison between time steps for a better normative and comparative explanation experience and a 'undo-redo' based activity history. Also, the user has no opportunity to seek into the original sensor data, so the concept misses enough transparency.

5.4 Improved concept based on analysis

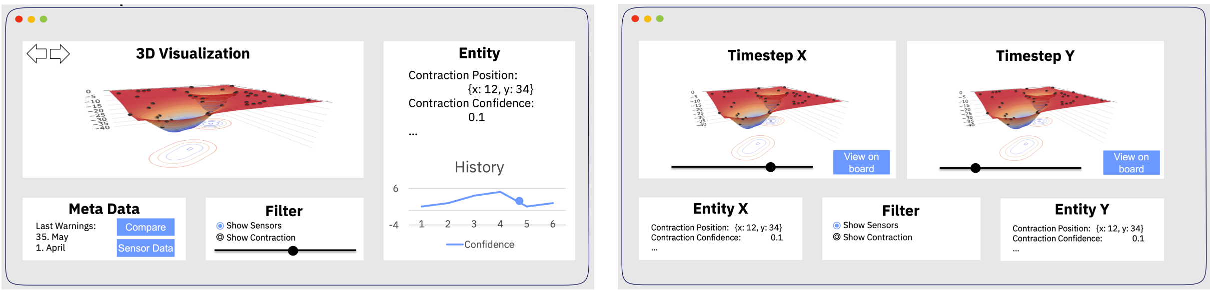

Figure 6: Improved Concept of the UI for the WDPS. On the left side is the primary interface shown, with more action buttons for the user than in the first concept. The right picture shows the comparison UI where the user has the ability to compare two specific timeframes next to each other. [49]

With the improved concept the user now has the ability to undo and redo his actions with two buttons on the top left corner. This way the user is able to apply changes to the visualization without loosing the current status. Another improvement wich comes with the undo and redo buttons is the capability of the system to be more comparable between different states accessable via undo and redo actions. The capability of the system to be more comparative is also improved with the included comparison mode, which can be seen on the right side of Figure 6. The comparison mode is accessible over a button in the meta field of the visualization. The mode has two swarm visualizations from selectable time steps so the user can compare these two. Under these two visualizations exist also two entity field with information of selectable entities. In the centre is the filter field, where the user can filter out irrelevant data. [50]

He is now also able to take a look into the original sensor data log over a button in the meta field. This way the system is more transparent with the data it consumes. [51]

5.5 Prototype of the 3D Visualization visualizing the swarm intelligence of the Water Damage Prevention System

[49] Improved Concept
[50] See Figure 6
[51] See Figure 6
[52] Prototype

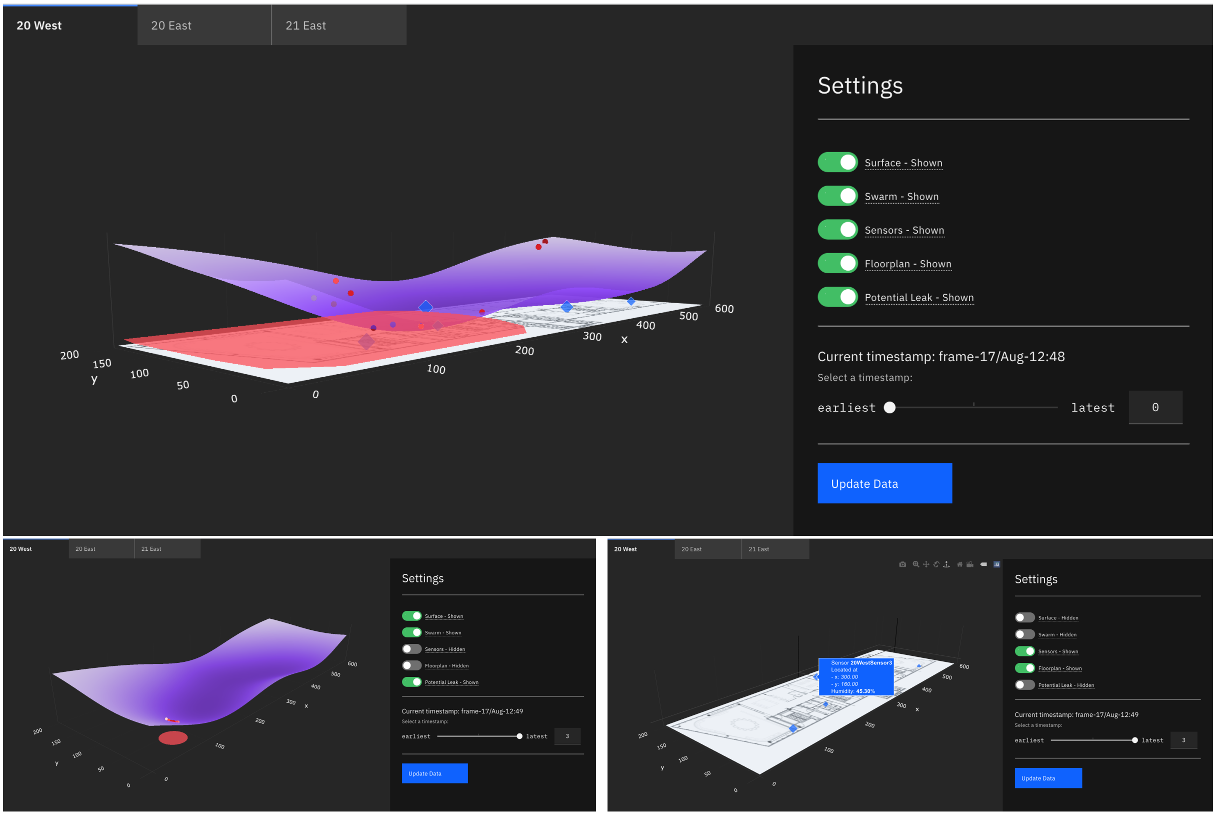

Figure 7: The prototype which was developed based on the concept above. The prototype misses other than the concepts an entity section. The meta and filter sections are merged together and the comparison part is also missing. But the 3D Visualization has tow more visible entities: the contraction circle and the floorplan. In the top image the visualization is shown with all entities visible at timestep 0. The left bottom picture shows the visualization with only the swarm and its environment surface visible for the third timestep. The right bottom picture shows also for the third timestep the floorplan with the sensors above while the mouse is hovering over a sensor and a small tooltip with useful information is shown. [52]

With some time problems it was not possible to implement the full improved concept into a prototype. It was only possible to include the most important component of the concepts; the 3D visualization. The prototype of the 3D visualization is made for the IBM Watson Center in Munich for the 20th and 21st floor of the eastern building and the 20th floor of the western building. The goal was to find the exact position of potential water leaks on a specific floor. Because of the specific use case the 3D visualization was expended with two more entities. The floorplan entity, which can be found below the surface gives the user more context where the scene takes action. The contraction entity is on top of the floorplan and gives the user visually a specific range where a potential leak could be. Also, the sensors moved from the top of the surface below the surface on top of the floorplan. With these additions it should be possible for the user to understand the context of the setting better.

The 3D visualization is made with Plotly.js. Plotly.js is an open source graphing library for data visualization and is based on d3.js and stack.gl.[53] With Plotly.js it is possible create many kinds of charts, 2D as well as 3D ones, in a declarative programming style.[54] Other UI elements are made with the IBM Carbon Design Component System (Carbon).[55] The theming and the colors of the UI are implemented also with Carbon. The decision to use Carbon was made because the Prototype is also used in an demonstration of IBM programms.

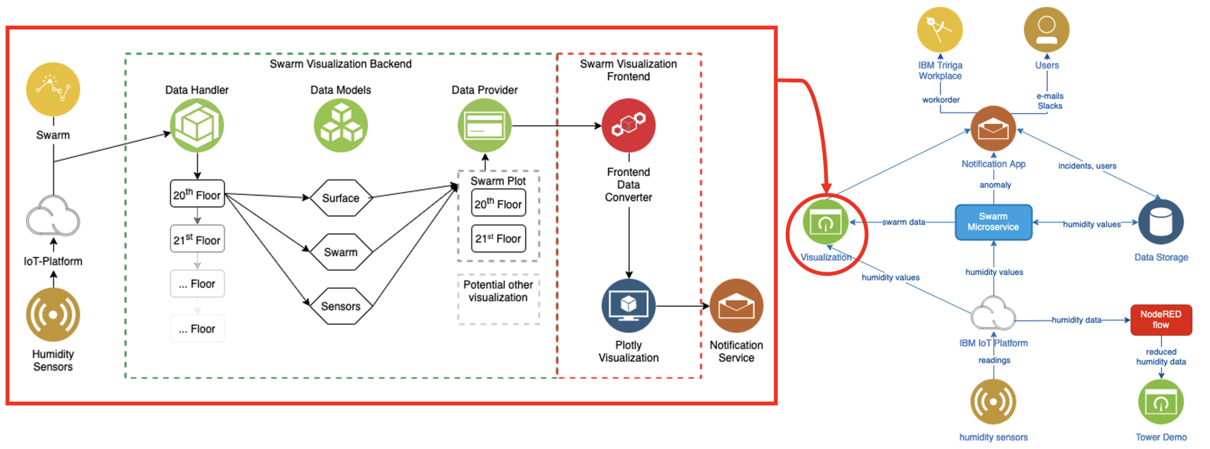

Figure 8: The service architecture of the WDPS on the right side with the data architecture specific for the visualization service (this prototype). [56]

The visualization is implemented as a microservice which serves in a larger architecture of the WDPS.8 The visualization microservice is splitted into a NestJS[57] backend and a VueJS[58] frontend which communicate over a SocketIO[59] connection. The basic principle of the communicate is a simple request and response pattern. The frontend requests after page load for visualization data from the backend. In response this data is then provided by a data provider in the backend. The provided data is then packed into the declarative PlotlyJS schema and rendered by the frontend. But before the data can be provided to the frontend, it must first be collected. The backend subscribes to the Sensors of the IoT Platform which sends the measured sensor values of each sensor. The swarm analysis microservice where the swarm is generated and its movements are calculated sends every 30 seconds all particle positions to the visualization microservice. The newest sensor and swarm data is called every 30 seconds from a data handler. This data handler updates then the data models which process and store the data for future use. The data provider then updates its own providing attributes based on the data models. This data can then be provided to the frontend. Behind everything runs a history handler which ensure that every model and provider saved the same amount of history steps. So there are two asynchron processes running: the data handler which updates the data models and providers

[53] Stackgl-community 2020; Bostock 2019
[54] Plotly 2020
[55] IBM 2020
[56] WDPS Architecture
[57] Mysliwiec 2020
[58] You 2020
[59] Socketio-community 2020

based on incoming data from the sensors and the swarm, and the data provider which hears for data requests by the frontend.[60]

An important part of the visualization is the contraction entity. This entity is a red circle on top of the floorplan which represents the radius where a potential leak could be. The radius of the red circle is calculated with the confidence of the swarm. If the confidence of the swarm is high, the circle radius is very small and points to a specific point on the floorplan. Else the circle radius is very large. The sensor are also represented on the visualization. Right over the floorplan floating some diamond shaped blue markers which represents the sensors. The larger the marker of a specific sensor is the more humidity was measured from the sensor. The user is also able to hover over the contraction point and each sensor to get more information about the specific entity. On hover over the contraction point a field appears where the current confidence and the coordinates in the virtual environment of the contraction point of the swarm is shown. On hover over a sensor the appeared field shows the exact coordinates in the virtual environment of the sensor as well as the measured humidity.

The swarm is represented by its particles which are represented by markers on the virtual surface entity. The nearer a particle is to the contraction point of the swarm, the darker is its colour. This helps to see which particle follow different timesteps. The virtual surface is on deeper points darker than on higher ones. This colouring indicates potential problems where the user should take attention on.

The prototype includes beneath the 3D visualization also a settings menu for the 3D visualization. The settings menu is spliced into three sections: a toggle section, a slider section and an update section. The user has in the toggle section the ability to show and hide each entity of the 3D visualization. Each toggle is also a modal with explanations what each entity represents and how they work. At the slider section the user has the ability to select a specific time frame of the swarm. There is also the date and the time of the current selected time frame shown. These toggles and the timeframe slider replace the filter section where the user should also be able to show and hide entities as well to select a specific timeframe. In the update section the user can click the "Update Data" button to request the newest data from the data provider service.

[60]See Figure 8

6 Conclusion

AI systems are very complex and hard to understand, what increases potential risk in use and decreases humans trust in them. Three aspects of increasing trust into AI are the increase of explainability, accountability and transparency of the system. Visualizations and UIs can potentially used to increase these factors. Therefore this work dealt with the topic of trustful AI and visualizations of such systems. The primary question was what are potential or necessary requirements of creating a good UI and UX for AI systems under the aspect to make them more trustful, in consideration of the practical use case of visualizing an AI system which uses a swarm intelligence and data from a sensor network.

In the first section of this thesis it was possible to define six key principles of trustful AI through a comparison between defined principles for trustful AI by leading IT companies. These are **fairness, explainability, transparency and accountability, security and privacy, social responsibility and technical availability**, where the principles **explainability, transparency and accountability** seemed to be easy applicable with creating a visualization for the AI system.

Potential necessary requirements of UI and UX for creating a trustful AI where found and collected in a structured literature review in the second section by analysing existing examples of visualizations of AI systems as well of sensor networks. Each requirement was formulated in a hypothesis which all where merged into one large hypothesis which connects requirements with its potential improving principles. These requirements are **normative and comparative explanations, quantified decision weights, irrelevance filter, fullfilling the seven tasks of visualizing sensor data and giving precise data insights**. [61]

Based on these requirements an artifact in form of a concept and a prototype was developed which is described and analysed in the third section. It was not possible to give an actual scientific proof of the hypothesis based on a survey or similar methods about the artifact. But it was possible to combine and fullfill all requirements in a concept without upcoming problems between the requirements. It was also possible to implement the 3D visualization from the concept into a Vue and NestJS framework environment, which is now used to explain the AI of the WDPS.

Therefore a final conclusion is that the necessary requirements for creating an UI and UX for an AI system are **normative and comparative explanations, quantified decision weights, irrelevance filter, fullfilling the seven tasks of visualizing sensor data and giving precise data insights**. These requirements helping fullfilling the trustful principles **explainability, transparency and accountability** and therefore such a visualization increases trust in an AI system.

[61] See Table 2

Bibliography

Bostock, M. (2019): d3.js. URL: https://d3js.org/.

Cai, C. J./Jongejan, J./Holbrook, J. (2019): The Effects of Example-Based Explanations in a Machine Learning Interface. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*, pp. 258–262. DOI: 10.1145/3301275.3302289. URL: https://doi.org/10.1145/3301275.3302289.

Chu, T./Yee, S. (2015): A visual Introduction to Machine Learning. URL: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/%20http://www.r2d3.us/visual-intro-to-machine-learning-part-2/ (visited on 08/02/2020).

Dobrev, D. (2004): A Definition of Artificial Intelligence. In: *Mathematica Balkanica, New Series* 19, pp. 67–74. arXiv: 1210.1568v1. URL: https://arxiv.org/abs/1210.1568.

EY (2018): How do you teach AI the value of trust? Tech. rep. EY, p. 8. URL: https://assets.ey.com/content/dam/ey-sites/ey-com/en%7B%5C_%7Dgl/topics/digital/ey-how-do-you-teach-ai-the-value-of-trust.pdf.

Google AI (2018): Responsible Development of AI. Tech. rep. Google, p. 20. URL: https://ai.google/static/documents/responsible-development-of-ai.pdf.

Holmström, J./Ketokivi, M./Hameri, A. P. (2009): Bridging practice and theory: A design science approach. In: *Decision Sciences* 40.1, pp. 65–87. ISSN: 00117315. DOI: 10.1111/j.1540-5915.2008.00221.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5915.2008.00221.x.

IBM (2020): Carbon Design System. URL: https://www.carbondesignsystem.com/.

IBM Research (2018): AI bias will explode. But only the unbiased AI will survive. URL: https://www.youtube.com/watch?v=q44XUZdCIMM%7B%5C&%7Dfeature=emb%7B%5C_%7Dlogo.

IEEE (2017): Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. The IEEE Global Initiative on Ethics of Autonomous/Intelligent Systems., p. 266. URL: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html%20https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead%7B%5C_%7Dv2.pdf.

Karaboga, D./Akay, B. (2009): A survey: algorithms simulating bee swarm intelligence. In: *Artificial Intelligence Review*, pp. 68–85. DOI: 10.1007/s10462-009-9127-4. URL: https://www.researchgate.net/publication/220638051.

Kennedy, J. (2006): "Swarm Intelligence". In: *Handbook of Nature-Inspired and Innovative Computing*. Boston: Kluwer Academic Publishers, pp. 187–219. DOI: 10.1007/0-387-27705-6_6. URL: http://link.springer.com/10.1007/0-387-27705-6%7B%5C_%7D6.

Metcalf, L./Askay, D. A./Rosenberg, L. B. (2019): Keeping Humans in the Loop: Pooling Knowledge through Artificial Swarm Intelligence to Improve Business Decision Making. In: *California Management Review* 61.4, pp. 84–109. URL: https://www.youtube.com/watch?v=jfoAYg-gk98%20https://doi.org/10.1177/0008125619862256.

Microsoft AI (2020): Responsible AI principles from Microsoft. URL: https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6 (visited on 07/07/2020).

Moor, J. H. (2003): The Turing Test - The Elusive Standard of Artificial Intelligence. Springer Science+Business Media Dordrecht. URL: https://books.google.de/books?id=7pFHBAAAQBAJ%7B%5C&%7Dprintsec=frontcover%7B%5C&%7Dredir%7B%5C_%7Desc=y%7B%5C#%7Dv=onepage%7B%5C&%7Dq%7B%5C&%7Df=false.

Mysliwiec, K. (2020): NestJS. URL: https://nestjs.com/.

Nguyen-Kim, M. T. (2020): Wissenschaftler Irren. URL: https://www.youtube.com/watch?v=DHyRaUeHcGY.

Plotly (2020): Plotly.js. URL: https://plotly.com/javascript/ (visited on 08/17/2020).

Polyakov, A. (2019): How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors) | by Alexander Polyakov | Towards Data Science. In: *towards data science*. URL: https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c.

Purdy, M./Daugherty, P. (2016): Why Artificial Intelligence is the future of growth. Tech. rep. Accenture. URL: https://www.accenture.com/t20170524T055435%7B%5C_%7D%7B%5C_%7Dw%7B%5C_%7D%7B%5C_%7D/ca-en/%7B%5C_%7Dacnmedia/PDF-52/Accenture-Why-AI-is-the-Future-of-Growth.pdf.

Ribeiro, M. T./Singh, S./Guestrin, C. (2016): "Why should i trust you?" Explaining the predictions of any classifier. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 13-17-Augu. Association for Computing Machinery, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778.

Rossi, F./Sekaran, A./Spohrer, J./Caruthers, R. (2019): Everyday Ethics for Artificial Intelligence. Tech. rep. IBM Corp., p. 48. URL: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

Scarr, S. W. (1989): "Protecting general intelligence: Constructs and consequences for interventions." In: *ntelligence: Measurement, theory, and public policy: Proceedings of a symposium in honor of Lloyd G. Humphreys*. Ed. by Robert L. Linn. University of Illinois Press, pp. 74–118. URL: https://books.google.de/books?hl=de%7B%5C&%7Dlr=%7B%5C&%7Did=ATTxTrkaCTUC%7B%5C&%7Doi=fnd%7B%5C&%7Dpg=PA74%7B%5C&%7Ddq=Protecting+General+Intelligence%7B%5C&%7Dots=rdDO-qGHUW%7B%5C&%7Dsig=wT5t2HOyESCh3RcPOXcPIQVhm54%7B%5C#%7Dv=onepage%7B%5C&%7Dq=Protecting%20General%20Intelligence%7B%5C&%7Df=false.

Socketio-community (2020): Socket.io.

Stackgl-community (2020): stack.gl. URL: http://stack.gl.

W Boers, E. J./Kosters, W. A./van der Putten, P./Poel, M./Kok, J. N. (2009): "Artificial Intelligence: Definition, Trends, Techniques and Cases". In: *Artificial Intelligence*. Ed. by Joost Nico Kok. Paris, France: Encyclopedia of Life Support Systems (EOLSS). Chap. 1. URL: http://www.eolss.net/sample-chapters/c15/e6-44.pdf%20http://www.eolss.net.

Will Douglas Heaven (2017): Predictive policing algorithms are racist. They need to be dismantled. URL: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/ (visited on 07/31/2020).

You, E. (2020): VueJS. URL: https://vuejs.org/.

Zankl, S. (2009): Visualizing Sensor Data. URL: https://www.medien.ifi.lmu.de/lehre/ws0809/hs/docs/zankl.pdf.

Declaration

I hereby affirm that I wrote my 1st Project Thesis with the topic: *UI / UX requirements for explainable and trustful AI systems* on my own and did not use any other sources and tools than those mentioned. I also assure that the submitted electronic version is identical to the printed version.

Ich versichere hiermit, dass ich meine 1st Project Thesis mit dem Thema: *UI / UX requirements for explainable and trustful AI systems* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Berlin, 27. August 2020
(Place, Date)

Tobias Hölzer
(Signature)