Does having an explanation improve trust and do users engage with the explanation? A case study on a movie recommendation system

Student Name: S.E. McFarlane

Supervisor Name: Dr L. Shi

Submitted as part of the degree of BSc Software Development for Business to the

Board of Examiners in the Department of Computer Sciences, Durham University

***Abstract* —**

Background - During the past decade, there has been an increase in the use of black-box models and with them a lack of understandability and therefore trust for the users. Explainable AI (XAI) has arisen to try and combat this issue and aims to create human understandable results from an artificial intelligence (AI) application. A fundamental part of XAI is the use of explanations of which there are many different varieties and they are said to help promote trust in a system for a user.

Aims - This project aims to answer the following research question, does having an explanation improve trust and do users engage with the explanation? This project will also aim to see user's preferences when it comes to explanations when given a choice from various types.

Method - Forty-one participants were asked to complete an explanation satisfaction questionnaire for three different explanations ranging from local to global scopes, static to dynamic explanations, visualizations to natural language explanations and explanations containing different levels of technical information. Two movie recommendation systems were created, with system A containing no explanations and system B having the highest scoring explanation from the previous round. To help overcome the cold-start problem and to give more accurate movie recommendations, users would be asked to rate five movies before being given their recommendations, as well as collaborative filtering being applied. Forty-six participants then took part in the A/B testing. Each participant was randomly allocated a system and asked to interact with it before completing a trust questionnaire whilst user behaviour on the system was also tracked.

Results - The static global explanation that was a combination of visualizations and natural language explanations scored the highest in the explanation satisfaction questionnaire. There was no statistically significant effect on trust, despite system B users obtaining a higher average trust score. System B caused increased engagement with the system through an increased number of clicks per user and a longer average time on the system.

Conclusions - Having explanations within an AI system does not lead to a significant increase in trust although trust is still a hard concept to measure. The inclusion of explanations does lead to an increase in overall engagement of the system but more research is still needed within the explainable AI field.

***Keywords* —** Explainable AI (XAI), explanations, trust, explainability, interpretable

I INTRODUCTION

Artificial intelligence is the domain for developing systems that have the characteristics we associate with human intelligence such as problem solving, learning and perception. It challenges technology to think and learn from past experiences. This domain is very broad and as such it not

only covers computing disciplines but also overlaps with other subject areas such as psychology, mechanical engineering, philosophy and mathematics. Artificial intelligence has many applications from search engines, recommendation systems, online banking and social media just to name a few.

Over the past decade, artificial intelligence has grown from being used only in high-level research, to now being used daily by millions of people across the globe. This is due to increasing research efforts as well as progression in the hardware now available to perform such computing on. Moore's Law has been instrumental in developing artificial intelligence and it states that the number of components that can be placed on a chip doubles every two years (Moore 1975). This has remained true to this day and has allowed chips to have enough memory and processing power to drive the artificial intelligence movement forward.

A Background

Explainable artificial intelligence (XAI) is a relatively new and emerging field within Computer Science. It aims to create human understandable results from an artificial intelligence (AI) application. During the past decade, there has been an increase in the use of black-box models and with them a lack of understandability and therefore trust for the users. Explainable AI has arisen to try and combat this issue. A black-box model is typically related to a deep neural network and is complicated and opaque meaning it is difficult for humans to understand. You can only see the inputs and outputs making it hard for humans to trust when used in critical and difficult tasks. Recently governments are starting to show more interest in XAI, especially with the European General Data Protection Regulation (GDPR) (High Level Independent Group on Artificial Intelligence 2019) being introduced, which has increased the urgency for XAI.

Explanations are used within explainable AI to promote transparency to make these systems more interpretable to the human user in the hope of promoting trust. There are many different types of explanations and each has a different domain and purpose.

As artificial intelligence moves into more critical and difficult fields such as autonomous vehicles (You et al. 2019), health care (Torres et al. 2018) and cyber-security (Parra et al. 2020), more questions are being asked as to how and why it makes a certain decision and to what level of accuracy it can make a decision. Explainable AI is at the forefront of this to help prove to users as well as experts in the relevant fields that AI applications can be used in these scenarios and not only excel but also be trusted with these vital and challenging decisions.

B Definitions

Many of the concepts and terminology within explainable AI are greatly contested within the literature, especially with the use of explainability and interpretability. Definitions are provided below to clarify what is meant throughout this paper.

Definition 1: Explainability - Notion of an explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans. (Guidotti et al. 2018)

Definition 2: Explanation - An answer to a why question. (Overton 2011)

Definition 3: Trustability - A measure of confidence, as humans, as end-users, in the intended working of a given model in dynamic real-world environments. (Das & Rad 2020)

Definition 4: Understandability - Denotes the characteristic of a model to make a human understand its function - how the model works - without any need for explaining its internal structure or the algorithmic means by which the model processes data internally. (Montavon et al. 2018)

Definition 5: Interpretability - Denotes the degree to which an observer can understand the cause of a decision. (Biran & Cotton 2017)

Definition 6: Transparency - Expresses enough to be human-understandable. (Das & Rad 2020)

C Project Objectives and Achievements

This project sets out to answer the research question, does having an explanation improve trust and do users engage with the explanation? On top of this, it aims to look at user's preferences when it comes to explanations of various types. To answer this research question, the project was broken down into objectives.

- **Minimum Objectives**
  - Build initial website including servers and database
  - Implement the movie recommendation case study into website
- **Intermediate Objectives**
  - Track user behaviour
  - Incorporate explanations
  - Interactive questions to improve recommendations
- **Advanced Objectives**
  - Implement an epsilon greedy algorithm into the recommendation algorithm
  - Incorporate user behaviour into recommendation algorithm

This project implemented the above objectives and has shown that the inclusion of explanations leads to increased engagement of the system, but there was no significant increase in trust. It has also shown that users do have a preference when it comes to explanations but again this doesn't lead to a significant increase in trust in the system.

II RELATED WORK

A Stakeholders in XAI

A constant debate within this field is who the stakeholders involved in explainable AI are. Some believe there are four groups, developers who build the applications, theorists who are concerned with AI theory and evolving it, ethicists who care about the fairness and transparency of systems and users who use the end application (Preece et al. 2018). The user groups can be broken down further into novices, those without any prior knowledge of the field, data experts in data analytics, and AI experts concerned with designing machine learning models (Mohseni et al. 2020). But the

one overarching concept that is widely adopted is that users should be the focus when designing and implementing explainable AI applications (Kirsch 2017) (Ribera & Lapedriza 2019).

We split individuals into different stakeholder groups as different explanations are needed depending on the user group (Ribeiro et al. 2016), as each group has different goals (Tomsett et al. 2018) and motivations (Sperrle et al. 2020). The prior knowledge of the different stakeholders affects the explanations we can give whilst still being interpretable to the relevant group. We aren't just concerned with the knowledge of users but also the knowledge of the AI system, this can be split into knowns and unknowns (Preece et al. 2018) and can affect the explanation we give to the different user groups. Accountability is a crucial factor in this field (Diakopoulos 2016) and can vary depending on the stakeholder. Ethicists are the ones that are most concerned with this however developers also play a key role as they ultimately implement accountability into the applications. Accountability is concerned with making each user group responsible for their actions. Interpretability can also fluctuate depending on stakeholders, as each will have different interpretability goals (Tomsett et al. 2018). Explainability in these scenarios is crucial for the stakeholders to reach their desired goals.

B Explanations

Explanations are fundamental to explainable AI. They are the main tool used to make black-box models interpretable. As seen in the definitions in the Introduction section B, explanations have a close link to why-questions. A whether-question is simply a question whose answer is yes or no. We will use Bromberger's definition of a why question as a whether-question that commences with a why (Bromberger 1966). Explanations can have type signatures which "distinguish between explanations with the same presupposition (i.e. explanandum) but different kinds of answers (i.e. different explanans type)" (Overton 2011). The goals of explanations are to facilitate learning, verify systems, comply with legislation such as General Data Protection Regulation (GDPR), improvement of the system (Samek et al. 2017), acceptance of the technology (Lim et al. 2009) and troubleshooting (Ribera & Lapedriza 2019).

While explanations are important many different types need to be considered. On a broad scale, there are local and global explanations that define the solution space of the explanation. Within these categories, we can also have pragmatic and non-pragmatic explanations which categorize whether an explanation is correct or simply just good (Kim 2018). We also have static or interactive explanations which are concerned with adapting or not to feedback from the customer (Arya et al. 2019). Layered explanations can also be used that incorporate traceability, justification, and assurance into the explanations (Preece et al. 2018). Another form of explanation is to explain the boundaries of the system, and therefore the limitations (Mueller & Klein 2011). Within each category, there are specialized explanations such as natural language explanations, visualizations, explanations by example, counterfactual examples and transparency-based explanations just to name a few. Lipton introduced the idea of post-hoc interpretability which is an approach to present useful information to the users (Lipton 2017). This includes some of the previously mentioned techniques such as natural language, visualizations and explanations by example. Post-hoc explanations focus on what the model can tell us rather than how the model works.

Within the literature, there is a consensus that explanations and in particular why-questions are contrastive (Miller 2019). Contrastive explanations explain relevant to another event and as such can also be placed in the counterfactual category. The process of designing explanations

can be taxing. Deciding what to explain and what not to can be seen as a challenge and as such we need to know what the triggers of explanation for the system are (Hoffman et al. 2018). Designing explanations is a co-adaptive process, it relies on both the explainer and the explainee. It is also seen as a continuous process to develop and maintain the trust in the system (Hoffman et al. 2018).

C Trust in XAI

Trust is a driving factor in explainable AI as it is the main motive for designing and implementing explainable AI tools and techniques. Each stakeholder group has different motives and goals but trust is integral to them all. To make a system transparent, trust in the system and/or explanation is needed (Das & Rad 2020). Explainable AI aims to increase trust and transparency to all users. To improve trust we have to show why a particular decision was made and thus explanations are of prime importance (Rossi 2018).

There are different levels of trust and with them, different scopes. We can trust in an explanation but that doesn't mean we have trust in the overall model (Sperrle et al. 2020). To gain the full trust of the user, they must trust at all levels and scopes involved. Explanations are usually used to develop trust in a system but if the explanations are poorly formed this can hinder the process. Trust differs from individual to individual and as such we must tailor the way we build trust to the different individuals and user groups. Different individuals will have justified and unjustified choices in trust or distrust, depending on their experiences (Hoffman et al. 2019). Trust is delicate, once lost it can be tough to re-build.

Trust can be a valuable metric for evaluating explainable AI and yet a very hard one to measure. Trusting a system is seen as an exploratory process and can vary from positive to negative (Hoffman et al. 2019). As it is exploratory and always changing this makes it difficult to evaluate. Past work has used trust measurement scales, but these can be very specific depending on the domain and context of the research.

To help improve trust, this project will give the user a choice of three different types of explanations, with the highest rated one implemented into system B of the movie recommendation system. This will be used to determine whether the inclusion of an explanation does promote trust for the user.

D Evaluation in XAI

Evaluation in explainable AI covers a lot of different aspects. There are evaluations of individual concepts such as explanations or trust as mentioned before, but then there are evaluations of a whole system as well. Different evaluation metrics can be used depending on the user group involved (Mohseni et al. 2020), the explanations used and the need for evaluation (Ribera & Lapedriza 2019). Some common measures include the user's mental model of the system, explanation usefulness and satisfaction, user trust and reliance, human-AI task performance and computational measures (Mohseni et al. 2020). These are all great metrics, but some are much harder to evaluate than others. User trust is a difficult metric to measure whereas computational measures are more straightforward to measure, as there is already a well established scale in place to refer to. Several of these metrics rely on a user and as such, they should be a key part of the evaluation process (Kirsch 2017) but this can present a challenge when evaluating as each user has different thoughts and feelings towards a given system.

Evaluation measures can be categorized into six classes: goodness of explanations, user satisfaction, user's understanding, curiosity, user's trust and reliance, and system performance (Hoffman et al. 2019). Many of these metrics rely on scales such as the explanation satisfaction scale (Hoffman et al. 2019) but these are challenging to validate. Checklists can also be used as a measure; the literature presents the explanation goodness checklist (Hoffman et al. 2019). While some scales and checklists may appear alike, they are intended for different audiences. The explanation satisfaction scale is concerned with the user's evaluation as the explanation goodness checklist is concerned with other researchers' evaluation.

While some prior work focuses on the evaluation of explanations and the system, other work focuses on evaluating the user interface resulting in human computer interaction being frequently commented on. Usability principles and how they can be used in the context of explainable AI are often considered. Norman proposes four design principles: user's conceptual model, visibility, mapping, and feedback (Norman 2013). While these were designed not with explainable AI in mind, each principle contains elements of explainability and are a good foundation to consider when building an explainable AI application.

To evaluate this project, the evaluation will be split up into different stages. An explanation satisfaction scale will be used to determine which explanation the users rate the highest, followed by a trust scale to see the level of trust a user has for the system. In addition to these scales, behaviour analysis will be conducted to judge the engagement levels of the users.

E Summary

Explanations are fundamental to promoting interpretability and explainability. As shown above there are many different categories and varieties that these can fall into. Prior work offers numerous explanations and describes the different scenarios and users that these might appeal to. We know that users play a key role but each individual user is different and brings with them different requirements and motivations and as such, this project will look at several different explanations in order to promote trust in the end system.

Many elements contribute to explainable AI. But within these elements, there is a pressing urgency to have the user at the centre when designing and implementing these applications. We build these applications to promote trust, but the users are the ones that must trust the system. When building the system in this project the users will be put at the forefront to ensure their needs are met and to help enhance interpretability within the system.

This project aims to build upon prior work by looking at the trust of users to see if this is affected by having an explanation as well as looking at what explanations users prefer. It will also show how users interact with explanations to see if they are used when placed in a system.

III SOLUTION

The solution for the research question is outlined below. There are two main stages, the first is a movie recommendation system and the second is related to the explanation itself and integrating it into the recommendation system.

A Overall Design

To test the research question, a movie recommendation system would be built and A/B testing applied. System A would have no explanations, as system B would have an explanation after

each movie for the user to use. This system would be implemented in the form of a website with a log-in system for repeat users. It would offer movie recommendations to the user based on their likes and dislikes as well as tracking user behaviour in the background.

B Movie Data

The movie data was sourced from the MovieLens Dataset (Harper & Konstan 2015). This provided one hundred thousand movies with information such as genre, runtime, and popularity as well as information on user's ratings such as vote count and vote average. The data was cleaned and only movies with thirty or more votes (the top 26%) were kept to leave one thousand, two hundred and seventy-eight movies remaining. All numeric values were normalized using Min-Max normalization.

The movies would be split into five groups to help with the recommendations, this was done through K-Means Clustering (Likas et al. 2003). An arbitrary number of five was selected for K as it was felt this would provide enough different groups for the recommendations to be based off. Budget, popularity, revenue, vote average and vote count were used to form the clusters.

Within each cluster, the movies were then ranked based on their distance to the centroid and popularity. The top twenty movies from each cluster were taken and put into a database leaving one hundred movies to be used in the movie recommendation system.

C Cold-start Problem

Recommendation systems work based on user's preferences and interests. A problem occurs when we get a new item or user as we have no prior knowledge of their preferences and interests, this is known as the cold-start problem (Park & Chu 2009). To overcome this, two techniques were used. When a new user signs up to the site, they would be asked to rate five movies, one from each movie cluster. They would be presented with three movies from each cluster and asked to rate one of them or none if they had never heard of any of them. This would allow us to collate information about their likes and dislikes to be used further down the line in the recommendation process.

Secondly, a common technique of collaborative filtering (Goldberg et al. 1992) was implemented. Collaborative filtering works by grouping users together with similar tastes to create a list of suggestions based on the other users. A table is used to store this information with each user having a row, and each movie having a column. The table is then populated with the user's ratings of each movie. This technique was adapted within this implementation and instead of using individual movies and the user's ratings of them, this was done off the movie clusters previously formed. So we had five columns within the table, one for each movie cluster.

As this was a brand new system with no prior user base, in order to populate the table for collaborative filtering the IMDb ratings (IMDb 2021) were used. This allowed us to access the ratings by demographic. From here rows were created in the table for females aged 18-29, males aged 18-29, females aged 30-44, males aged 30-44, females aged 45+ and males aged 45+. This along with the initial five movie ratings the user had been asked to do would be used to help form the recommendation.

D Recommendation Algorithm

In order to give the user movie recommendations a recommendation algorithm needed to be

created. The previous knowledge of the collaborative filtering table based on the IMDb ratings would be used as well as the initial user's ratings for the five movies, one from each cluster. First of all the collaborative filtering table would be updated to reflect the user's preferences based on their initial ratings. If they had not given a rating then the baseline of the IMDb demographic rating would be used. From here the highest ranked movie cluster for that user would be sent as the recommendation. A recommendation would be sent three movies at a time and these would be randomly selected from within the highest ranked movie cluster.

A common problem in recommendation systems is the exploration versus exploitation problem (Bondu et al. 2010). It is a problem whereby you want to maximize the items you know the user likes whilst still exploring and gaining knowledge on new items. To balance exploration versus exploitation an epsilon greedy algorithm was implemented. The first two recommendations would always be from the highest ranked movie cluster for that user, but the third movie would have an epsilon greedy algorithm attached. The highest ranked movie cluster would be shown with probability e and a random movie would be shown with probability $1 - e$. For this implementation e was chosen to be 0.3. This addition allowed the user to still see movies that were not naturally the best matched but that they could still enjoy, helping to balance the exploration versus exploitation problem.

To improve this algorithm further, buttons were added to the website below each movie recommendation to ask if the user thought this was a good recommendation for them. They could press either 'Yes' or 'No' and this would update the collaborative filtering table accordingly to further improve the recommendations being shown to the user.

E Explanations

As explanations are a co-adaptive process, the users would be asked to rate three different explanations with the most popular one being implemented into the final system. Each explanation was created with different concepts in mind to provide the largest difference possible between them.

Figure 1 shows explanation A which provided a global explanation of how the whole recommendation system worked. It combined a natural language explanation with visualization techniques to form a flowchart and mind map of the process. It simplified down the continuous process of updating the collaborative filtering table and the recommendation algorithm. This explanation focused on how the system works rather than what the system can tell the user.

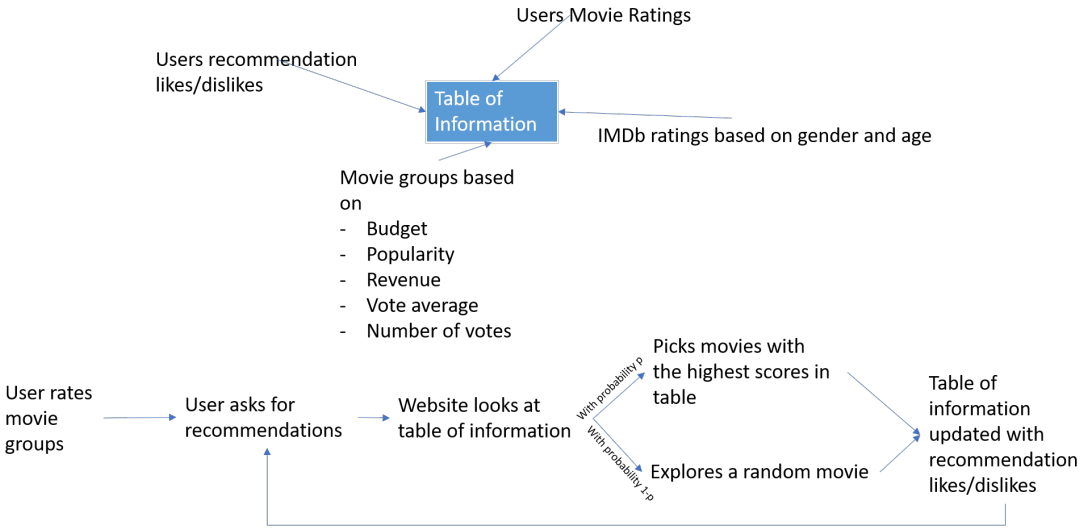

Figure 1: Explanation A

Figure 2 shows explanation B which was a dynamic local explanation for each individual movie recommendation. It would be personalised to each user and rely on visualization techniques. It showed the user the main data that is used to provide them with the recommendation and would dynamically be updated based on their perceived likes and dislikes from the collaborative filtering table. It included a lower degree of technical information compared to explanation A but was dynamic and personalised. It relied on a simple mathematical equation in the visualization to build upon the user's previous knowledge to help construct an understanding of how the system works.

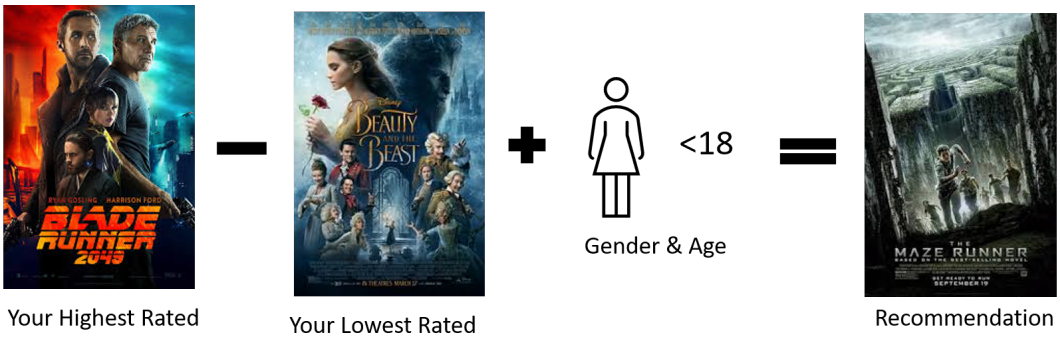

Figure 2: Explanation B

Figure 3 shows Explanation C, a global explanation in the form of a hierarchical tree structure. At the top of the structure is the recommendation itself and the leaves are the data that gets provided to the recommendation algorithm. Again this is a visualization explanation and is static, it would not adapt to any changes in the data or system. Similar to explanation B it did not contain any technical information, instead just showed what information is used to make the recommendation for the user.

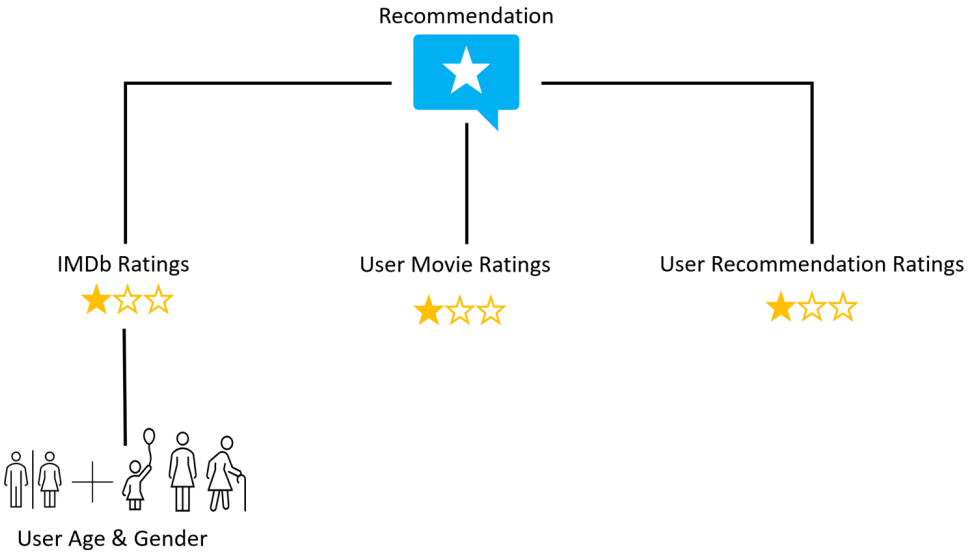

Figure 3: Explanation C

Table 1 shows a summary of the key differences between the explanations. This includes their scopes, explanation types, technical information included, personalization, and whether they are static or dynamic. The three explanations, A, B & C cover a wide range of explanations so the user can easily differentiate between them and so we can see where the user's preferences lie.

Table 1: SUMMARY OF EXPLANATIONS

Explanation	**Scope**	**Dynamic or Static**	**Technical Information Included**	**Personalised**	**Explanation**
A	Global	Static	Yes	No	Natural Language & Visualization
B	Local	Dynamic	No	Yes	Visualization
C	Global	Static	No	No	Visualization

F Tracking User Behaviour

Within the systems, user behaviour would be tracked in the form of buttons pressed. Every time a button was pressed a new data entry would be created stating the user, button ID and timestamp. These would be collated into a dedicated table in the database. To increase the information we could obtain from tracking, two new buttons were created, similar to the yes and no buttons for the recommendations. The users would be asked if they trusted the recommendation and presented with a 'Yes' or 'No' button to which they could respond. The log-in system would provide a list of users as well as allowing individuals to re-visit the site.

When implementing the tracking behaviour into the website, a problem was established. As an in-memory database of SQLite had been chosen it made it difficult to access the user data. To overcome this buttons were hidden within the web page itself that could be accessed from the console and provide all the information on users and their behaviour from the required databases.

G Verification and Validation

To help verify and validate the system, a prototype was created to be tested on people. A prototype with the above implementations was presented to two individuals to gain insight into their thoughts on the user interface as well as system features. This was done to reduce the risk for the final product and to validate the system as a movie recommendation system. Validity checks were completed to ensure all functions to support a user's needs were included and completeness checks were also performed to ensure all functions required by the user were included. As trust scores would be collected using a questionnaire and tracking user behaviour, this would assist in validating the final results. The trust results would come from two different sources, a subjective source in the questionnaire and an objective source in the user behaviour tracking.

H Testing and Deployment

To test the system each feature was subject to unit tests before being subject to integration tests with the existing system. These were done with a top-down approach so existing features could be built upon. This testing process was crucial to risk reduction in the final end system. Prototyping was also used to gain user insights as the user was at the centre of the design for the system. Once all tests had been completed, the system was copied and the explanations removed, this created system A. System B was left as is, with the explanations included. The two systems were hosted using Heroku which also allowed for continuous delivery and testing.

I Tools & Software Used

JavaScript was chosen as the primary programming language, with Node.js packages used to support this. Express was used to build the server and jQuery used to communicate between the front end and the back end of the website. Bootstrap was used to aid in the design of the front-end and allowed for a professional finish.

A relational database was needed for this website and SQLite was chosen for the database integration as it gave all the features of SQL but allowed for a lightweight implementation as this system would not be seeing a lot of traffic. Within the SQLite database itself, there were eight tables, one for each movie cluster, one for all movie entries, one for tracking user behaviour, and one for users.

J Final Implementation

The final end-system brought together everything that has previously been mentioned to form two movie recommendation systems: system A and system B. Screenshots of the final products can be seen in Figures 4, 5 & 6.

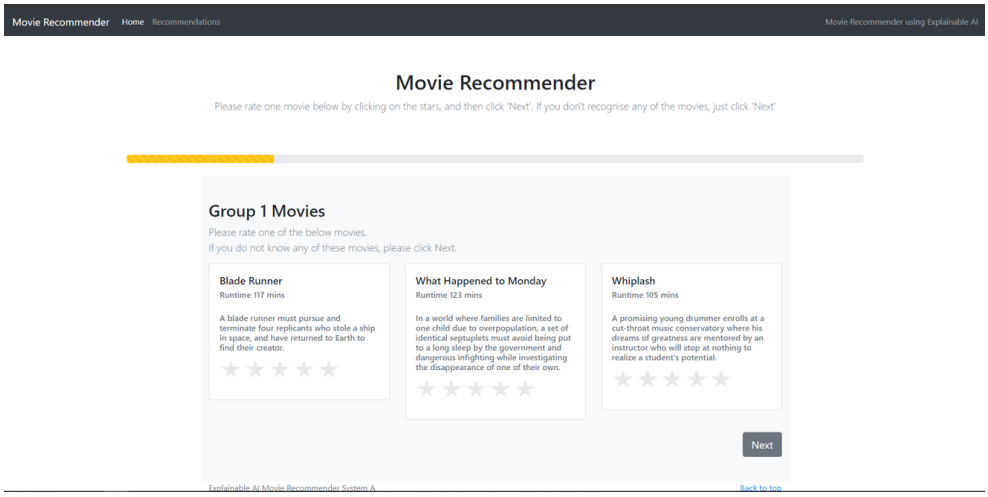

Figure 4: Users being presented with movies to rate for the movie recommendations

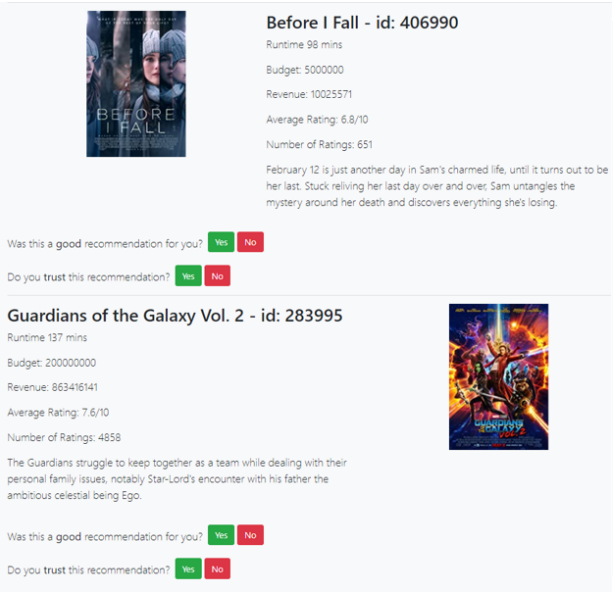

Figure 5: System A recommendations

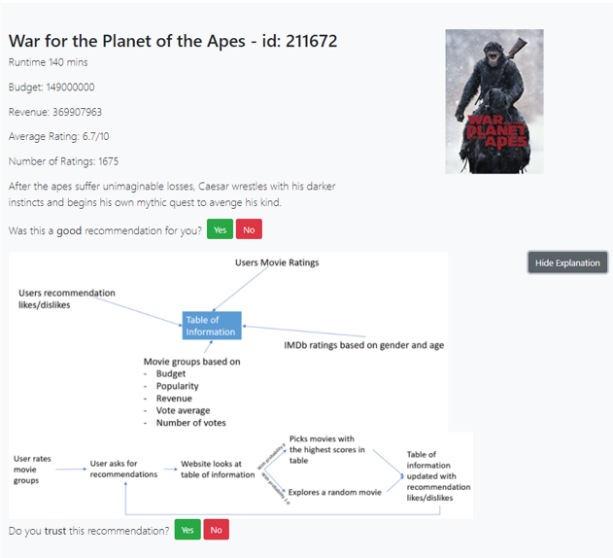

Figure 6: System B recommendation

IV RESULTS

This section presents the results of the explanations and trust of the movie recommendation systems.

A Explanation Results

Forty-one participants volunteered to take part in this study, including thirty females and eleven males. Table 2 shows the demographics of the participants.

Participants were informed of the protocol prior to providing written consent to participate. Each participant was given the three explanations (explanation A, B & C) and asked to complete an Explanation Satisfaction Scale (Hoffman et al. 2019) for each one. This scale was chosen

Table 2: SUMMARY OF PARTICIPANT DEMOGRAPHICS FOR EXPLANATION SATISFACTION SCORES

Ages	**No. of Males**	**No. of Females**
18-29	8	14
30-44	0	4
45+	3	12

as it has been developed to evaluate explanations by users, it has also proved to be valid in the XAI context through content validity and discriminant validity (Hoffman et al. 2019). Content validity shows that the scale items are meaningful and relevant to the domain and consistent with associated theories and frameworks, as judged by domain experts. Discriminant validity is used to demonstrate that the scale can differentiate between good and poor explanations. The average explanation satisfaction score for each explanation can be seen in Figure 7.

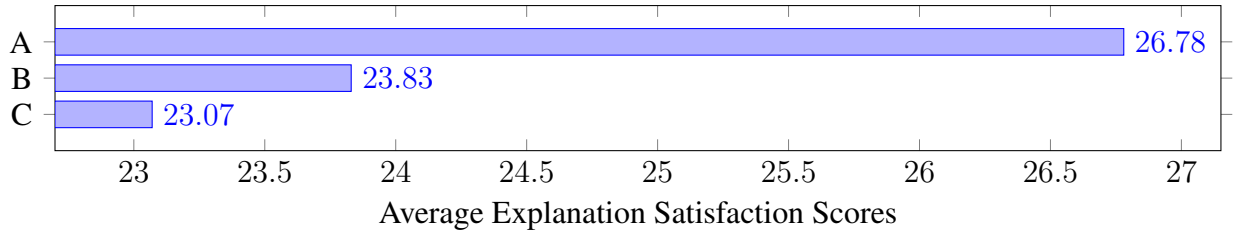

Figure 7: Graph showing the average explanation satisfaction scores for explanations A, B & C

Performing a Kruskal-Wallis test (Kruskal & Wallis 1952) on the explanation satisfaction scores gave a test statistic of $\chi^2(2) = 5.690$ and p-value of $p = 0.058$, showing that the null hypothesis could not be rejected and that the differences between the medians are not statistically significant, although the p-vale is remarkably close to being significant. When we extend this further to independent t-tests to focus on the differences between just two sets of scores we get the results shown in Table 3.

Table 3: INDEPENDENT T-TEST BETWEEN EXPLANATION SATISFACTION SCORES

Independent t-test between	**Test Statistic**	**P-Value**
Explanation A & B Satisfaction Scores	2.012	0.048
Explanation A & C Satisfaction Scores	2.448	0.017
Explanation B & C Satisfaction scores	0.454	0.651

From this we can see that explanation A has a statistically significant increased explanation satisfaction score when compared to explanations B & C. While the differences in scores between explanations B & C are not statistically significant, as the p-value is too large, stating that the null hypothesis cannot be rejected. As explanation A had the highest mean average score and is the most significantly different when compared to the other explanations, this was chosen to be implemented into system B of the movie recommendation application.

B Trust Results

Forty-six participants volunteered to take part in the second stage of this project, consisting of twenty-three females and twenty-three males. The demographics of the participants can be seen in Tables 4 & 5.

Table 4: SUMMARY OF PARTICIPANT DEMOGRAPHICS FOR SYSTEM A

Ages	No. of Males	No. of Females
18-29	9	10
30-44	0	0
45+	3	1

Table 5: SUMMARY OF PARTICIPANT DEMOGRAPHICS FOR SYSTEM B

Ages	No. of Males	No. of Females
18-29	8	8
30-44	1	0
45+	2	4

Each participant was informed of the protocol before providing written consent. From here, each participant was randomly allocated either system A or system B to conform with A/B testing. The two systems were identical except for system B containing explanation A (Figure 1) for each movie recommendation. This was done to isolate the explanations to see the effect the inclusion of the explanations would have on the user's trust. The participant was instructed to interact with the given system for ten minutes before being asked to conduct a trust questionnaire. The trust questionnaire given to the participants was recommended for use in XAI (Hoffman et al. 2019) and items within this scale have been adapted from the Cahour-Forzy Scale (Cahour & Forzy 2009), Jian, et al Scale (Jian et al. 2000), Schaefer Scale (Schaefer 2013) and the Madsen-Gregor scale (Madsen & Gregor 2000). This trust scale can be assumed to be reliable as it has considerable overlap with the Jian, et al (2000) scale and semantic similarity to items in the Madsen-Gregor Scale (Hoffman et al. 2019), both of which have been shown empirically to be highly reliable. We can also assume it has content validity due to the overlap with items in most of the already existing scales. The average trust scores for each system can be seen in Figure 8.

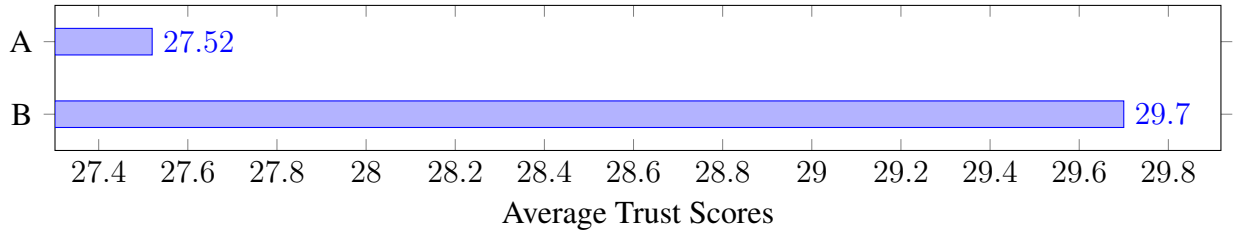

Figure 8: Graph showing the average trust scores for systems A & B

Performing an independent t-test on the trust scores gave a test statistic of $t(22) = 1.741$ and a p-value of $p = 0.089$. There was no statistically significant effect on trust, despite system B users obtaining a higher average trust score. Looking at the data from the user behaviour tracking in Table 6 shows a different picture. The explanation buttons were pressed on average five times per user, showing that the users did interact with the explanation and this did lead to an increase in time on the website.

Table 6: SUMMARY OF USER BEHAVIOUR

	System A	**System B**
Avg No. of clicks per user	17.57	23.65
Avg Time on website per User (mins)	4.66	12.81
No. of 'Yes' trust buttons pressed	47	49
No. of 'No' trust buttons pressed	50	2
Avg No. of explanation buttons pressed per user	Na	5

Having the explanations in system B caused increased engagement with the system, in the form of the number of clicks and time on the website. 48% of trust buttons pressed in system A stated they did trust the recommendation compared to 96% in system B although this did not lead to significantly higher trust scores in the questionnaire.

C Results by Demographic

The explanation satisfaction scores by demographic can be seen in Table 7. This demonstrates that on average males rated each explanation higher than their female counterparts. The 18-29-year-old age bracket rated each explanation higher than the overall combined average and higher than the 45+ age bracket which was always below the overall combined average. However, every demographic still found explanation A to be the best followed by explanation B and then C, with the exception of the 30-44-year-old's who rated explanation C higher than B.

Table 7: SUMMARY OF EXPLANATION SATISFACTION SCORES BY DEMOGRAPHIC

Avg Explanation Satisfaction Scores	**Explanation A**	**Explanation B**	**Explanation C**
Overall	26.78	23.83	23.07
Males	30.18	24.36	23.18
Females	25.53	23.63	23.03
18-29-year-old's	27.09	24.54	23.23
30-44-year-old's	29.75	22.25	23.50
45+ year-old's	25.53	23.20	22.73

The trust scores broken down by demographic can be seen in Table 8, which shows that not one particular demographic had overall higher trust scores with them all being varied depending on gender, age and system. This shows that a particular participant is not more likely to trust a system based on their age, gender, or the system they were randomly allocated.

Table 8: SUMMARY OF TRUST SCORES BY DEMOGRAPHIC

Avg Trust Scores	**System A**	**System B**
Overall	27.52	29.70
Males	25.58	31.00
Females	29.64	28.50
18-29-year-old's	27.05	29.81
45+ year-old's	29.75	29.00

V EVALUATION

This section evaluates the strengths and weaknesses of this solution using the results from the previous section.

A Strengths and Limitations

The average explanation satisfaction scores from Figure 7, show explanation A having a significantly higher score, which was reinforced by the independent t-tests in Table 3. This proves it was the right explanation to be implemented into system B. While this explanation did increase the average trust scores as seen in Figure 8, it was not enough to be statistically significant. This could show that what users perceive to be the best explanation for a given scenario and domain is in fact not the best in terms of promoting trust. Users themselves may not be confident or have enough knowledge to pick an explanation that will help them gain trust in a system.

The explanation satisfaction scores by demographics shown in Table 7, demonstrated that males preferred the explanations including the implemented explanation A which did lead to a higher trust score in system B than the females as seen in Table 8. However, both demographics of females and 45+ year-old's gave explanation A the highest explanation satisfaction score but this still led to higher trust scores in system A over system B. Again this could show that users don't know what explanation is best for them, with their believed favourite not performing the best when it comes to promoting trust.

Participants were asked to interact with the system for a minimum of ten minutes which was clearly stated in the protocol. However, the average time on the website for system A users was considerably less as shown in Table 6. This could have reflected badly in the trust scores as users did not have enough chance to engage with the system and decide if they trusted it or not. Given sufficient time with the system, the participants could have had increased or decreased levels of trust.

Trust as a concept is very subjective and as such can vary from individual to individual as well as change over time. As stated in the literature, trust is a hard concept to measure as no predefined scale exists. This project used not only a subjective trust scale but also the objective user behaviour within the system itself to gauge the user's trust and help to validate the trust of the user. While the trust questionnaire did not yield any significant result for increased trust, the user behaviour did, suggesting that trust questionnaires are perhaps not the best way forward when measuring a user's trust in a system. While the average trust scores from the questionnaires did not produce a statistically significant increase, the number of 'No' trust buttons pressed was considerably lower for the system B users. This shows us that the objective user tracking can provide insight that the subjective questionnaires can't and that both should be used in conjunction to provide the best possible insight.

The demographics presented in Tables 4 & 5, show the majority of users were from the age bracket of 18-29-year-old's, this generation has grown up with technology, unlike the other age brackets and means they are far more familiar with it. This could have skewed the results either way as they made up 76% of the total participants. But as previously mentioned trust is subjective so even within this age bracket the trust scores varied so may not have had a lasting implication on the results. Nonetheless, the 18-29 age bracket did rate each individual explanation higher than the overall combined average as seen in Table 7 which did lead to them being the only age demographic to have higher trust scores in system B than system A. This could illustrate that the

higher the explanation satisfaction scores are, the more likely that demographic is to trust in the system.

This project and especially the results shown in Table 6 have shown users do engage with explanations if they are included in an artificial intelligence system. It led to increased time on the website and an increase in the average clicks seen per user when compared with system A users. In this project, that may not have led to a significant increase in trust, but it does not mean that in the future with a different scenario and explanation, a different outcome could be observed.

This project was organised and broken down into objectives to help achieve the aim of answering the research question: 'Does having an explanation improve trust and do users engage with the explanation?'. The use of objectives assisted in separating the project out and making sure the end goal was achieved, which ultimately led to a successful and complete project. A significant proportion of the project was taken up in finding and guiding participants in the questionnaires, however this had been well accounted for at the beginning of the project so did not harm any other aspect.

B Improvements

If this project was to be duplicated again, some small adjustments would be made. Firstly every participant would be asked to interact and complete a trust questionnaire for both system A and system B to give a baseline trust score for the system from which to judge if having the explanation increased from their baseline score. Secondly, a more varied participant demographic would try to be obtained to have a broader overview of the project as a whole and to see if different age brackets and genders led to different levels of trust in the system.

Explanations A, B & C were designed specifically for this implementation, so no prior models were used to generate these explanations. As such it would be hard to replicate these explanations into another domain or project as they were so specific. In the future, it would be more beneficial to implement explanations from a model so others could then replicate and learn from it further.

The log-in system whilst completed successfully added limited value in this scenario. As the participants had only been instructed to use the system for a minimum of ten minutes, no participant re-visited the system. Therefore making the log-in system redundant. It would have been more constructive to use the time used to implement the log-in functionality for implementing a more accurate recommendation algorithm.

VI CONCLUSIONS

All minimum, intermediate and advanced objectives were met successfully with a movie recommendation system implemented. This included the use of collaborative filtering and interactive questions to overcome the cold-start problem and an epsilon greedy algorithm implemented into the recommendations to balance exploration vs exploitation.

This project has shown that including explanations in an artificial intelligence system helps to increase overall engagement in the system. But it has also shown that the inclusion of such explanations does not necessarily increase the overall trust in the system, even if the users have engaged with the explanation in question. There is not a singular demographic which is more likely to trust a system but more research is needed to see if a certain type of explanation is more suited to a specific age or gender. One potential research area is to see if a user's perceived

favourite explanation does lead to the biggest increase in trust of a system or if another type that is not so highly rated by the user does in fact increase their trust in the system more.

Trust is still a very hard concept to measure and subjectivity can cause problems when trying to evaluate explainable AI systems. More research is needed to accurately measure trust in systems which can then be used in the explainable AI domain.

Explainable AI, although still a fairly new concept within computer science, is only going to grow more traction. Artificial intelligence systems have grown in popularity and are being used in more critical and demanding environments such as banking, security and health care settings. As these applications get introduced into these fields, more questions are being asked about their trustworthiness and reliability. Explainable AI is needed to reassure and prove that these systems are making the correct decisions and therefore increased research efforts are needed to aid in this process.

References

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D. & Zhang, Y. (2019), 'One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques', *arXiv* .

Biran, O. & Cotton, C. (2017), Explanation and justification in machine learning: A survey. unpublished.

Bondu, A., Lemaire, V. & Boullé, M. (2010), Exploration vs. exploitation in active learning : A bayesian approach, *in* 'The 2010 International Joint Conference on Neural Networks (IJCNN)', pp. 1–7.

Bromberger, S. (1966), Why-questions, *in* R. G. Colodny, ed., 'Mind and Cosmos – Essays in Contemporary Science and Philosophy', University of Pittsburgh Press, pp. 86–111.

Cahour, B. & Forzy, J.-F. (2009), 'Does projection into use improve trust and exploration? an example with a cruise control system', *Safety Science* **47**, 1260–1270.

Das, A. & Rad, P. (2020), 'Opportunities and challenges in explainable artificial intelligence (xai): A survey', *arXiv* .

Diakopoulos, N. (2016), 'Accountability in algorithmic decision making', *Commun. ACM* **59**(2), 56–62.

Goldberg, D., Nichols, D., Oki, B. M. & Terry, D. (1992), 'Using collaborative filtering to weave an information tapestry', *Commun. ACM* **35**(12), 61–70.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. (2018), 'A survey of methods for explaining black box models', *ACM Comput. Surv.* **51**(5).

Harper, F. M. & Konstan, J. A. (2015), 'The movielens datasets: History and context', *ACM Trans. Interact. Intell. Syst.* **5**(4).

High Level Independent Group on Artificial Intelligence (2019), *European Comm.* .

Hoffman, R., Klein, G. & Mueller, S. (2018), 'Explaining explanation for "explainable AI"', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **62**, 197–201.

Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. (2019), 'Metrics for explainable AI: Challenges and prospects', *arXiv* .

IMDb (2021), 'IMDb: Ratings, reviews, and where to watch the best movies i& tv'. (accessed: 25.01.2021).
URL: *https://www.imdb.com/*

Jian, J.-Y., Bisantz, A. & Drury, C. (2000), 'Foundations for an empirically determined scale of trust in automated systems', *International Journal of Cognitive Ergonomics* **4**, 53–71.

Kim, T. W. (2018), 'Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test', *arXiv* .

Kirsch, A. (2017), Explain to whom? Putting the user in the center of explainable AI, *in* 'Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017)', Bari, Italy.

Kruskal, W. H. & Wallis, W. A. (1952), 'Use of ranks in one-criterion variance analysis', *Journal of the American Statistical Association* **47**(260), 583–621.

Likas, A., Vlassis, N. & Verbeek, J. J. (2003), 'The global k-means clustering algorithm', *Pattern Recognition* **36**, 451–461.

Lim, B. Y., Dey, A. K. & Avrahami, D. (2009), Why and why not explanations improve the intelligibility of context-aware intelligent systems, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', CHI '09, Association for Computing Machinery, New York, NY, USA, p. 2119–2128.

Lipton, Z. C. (2017), 'The mythos of model interpretability', *arXiv* .

Madsen, M. & Gregor, S. (2000), Measuring human-computer trust, *in* 'Proceedings of the 11 th Australasian Conference on Information Systems'.

Miller, T. (2019), 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence* **267**, 1–38.

Mohseni, S., Zarei, N. & Ragan, E. D. (2020), 'A multidisciplinary survey and framework for design and evaluation of explainable ai systems', *arXiv* .

Montavon, G., Samek, W. & Muller, K.-R. (2018), 'Methods for interpreting and understanding deep neural networks', *Digital Signal Processing* **73**, 1–15.

Moore, G. (1975), Progress in digital integrated electronics, *in* 'Technical Digest of the International Electron Devices Meeting', IEEE Press, p. 13.

Mueller, S. T. & Klein, G. (2011), 'Improving users' mental models of intelligent software tools', *IEEE Intelligent Systems* **26**(2), 77–83.

Norman, D. A. (2013), *The design of everday things*, revised edn, Basic Books, Inc.

Overton, J. A. (2011), Scientific explanation and computation, *in* 'Proceedings of the 6th International ExaCt workshop', pp. 41–50.

Park, S.-T. & Chu, W. (2009), Pairwise preference regression for cold-start recommendation, *in* 'Proceedings of the Third ACM Conference on Recommender Systems', RecSys '09, Association for Computing Machinery, New York, NY, USA, p. 21–28.

Parra, G., Rad, P., Choo, K.-K. R. & Beebe, N. (2020), 'Detecting internet of things attacks using distributed deep learning', *Journal of Network and Computer Applications* **163**, 102662.

Preece, A., Harborne, D., Braines, D., Tomsett, R. & Chakraborty, S. (2018), 'Stakeholders in explainable AI', *arXiv* .

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016), '"why should I trust you?": Explaining the predictions of any classifier', *arXiv* .

Ribera, M. & Lapedriza, A. (2019), Can we do better explanations? a proposal of user-centered explainable ai, *in* 'CEUR Workshop Proceedings', CEUR Workshop Proceedings.

Rossi, F. (2018), 'Building trust in artificial intelligence', *Journal of International Affairs* **72**, 127.

Samek, W., Wiegand, T. & Müller, K.-R. (2017), 'Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models', *arXiv* .

Schaefer, K. E. (2013), 'The perception and measurement of human-robot trust', *Doctoral dissertation, University of Central Florida Orlando, Florida* .

Sperrle, F., El-Assady, M., Guo, G., Chau, D. H., Endert, A. & Keim, D. (2020), 'Should we trust (x)AI? design dimensions for structured experimental evaluations', *arXiv* .

Tomsett, R., Braines, D., Harborne, D., Preece, A. & Chakraborty, S. (2018), 'Interpretable to whom? a role-based model for analyzing interpretable machine learning systems', *arXiv* .

Torres, A., Yan, H., Aboutalebi, A., Das, A., Duan, L. & Rad, P. (2018), *Patient Facial Emotion Recognition and Sentiment Analysis Using Secure Cloud With Hardware Acceleration*, pp. 61–89.

You, C., Lu, J., Filev, D. & Tsiotras, P. (2019), 'Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning', *Robotics and Autonomous Systems* **114**, 1–18.