

Identifying critical design elements for increasing trust in computers through co-designing XAI for a mobile banking app

Alex Blandin

953274

Submitted to Swansea University in partial fulfilment
of the requirements for the Degree of Master of Science



Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

27th October 2021

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed  (candidate)

Date 27th October 2021

Statement 1


This work is the result of my own independent study/investigations, except where otherwise stated. Other sources are clearly acknowledged by giving explicit references. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure of this work and the degree examination as a whole.

Signed  (candidate)

Date 27th October 2021

Statement 2

I hereby give my consent for my work, if accepted, to be archived and available for reference use, and for the title and summary to be made available to outside organisations.

Signed  (candidate)

Date 27th October 2021

I would like to dedicate this work to my parents, who made all this possible.

Abstract

Explainable Artificial Intelligence (XAI) has established itself as a central tool for developing advanced AI systems in even high-stakes domains, as techniques of explanation have underpinned an increased trust in their deployment by experts. While there is excellent research and ongoing developments in this area, it was felt that there was insufficient understanding of how to apply XAI properly when the explanations were not designed for experts.

To develop this understanding, particularly in the area of trust where XAI markets itself, this project used a human-centred approach to develop substantiated design concepts and identify practical design elements. This was achieved by a participatory design process, producing XAI for select scenarios from a mobile banking app; stylised to resemble the Starling Bank mobile app, to improve believability for participants.

The studies involved demonstrated that interpretable “glass box” models which produce written, textual outputs in a concise, plain language manner, was the desirable explanation method for users. These leveraged existing financial literacy and natural language skills, rather than requiring the user learn how to read a more graphical explanation. The applicability of such explanations as supplementary diagrams was established, however participants recognised that these were not always desired, particularly in time-critical contexts, where the textual designs were considered preferable and sufficient.

From our findings, we establish a few heuristic guidelines that can be considered for future research, particularly in XAI for banking. These findings demonstrate that there are indeed aspects where human-centred XAI approaches similar to the one pursued in this project can reveal the contextual nuance of an XAI deployment, which can provide valuable insight into what qualities are essential to the XAI used, enabling prioritisation and designs that contrasts in areas with broader XAI guidelines.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Overview	2
2	Background	3
2.1	Relevant Literature	4
3	Methodology	8
3.1	Focus Groups	9
3.2	Thematic Analysis	14
3.3	XAI Design	14
3.4	Evaluation	16
3.5	Limitations	18
4	Results, Designs, and Findings	19
4.1	Thematic Analysis of Focus Groups	19
4.2	XAI Designs	31
4.3	Findings from Evaluation Session	45
5	Conclusions	49
5.1	Contributions	49
5.2	Future Work	50
	Bibliography	51
	Appendices	55

A	Extended Questions for Focus Groups	56
A.1	Methods of XAI	56
A.2	Experience with XAI in a Financial Context	57

Chapter 1

Introduction

This dissertation documents a project that began with the realisation that, currently, there is little understanding of what makes an appropriate Explainable Artificial Intelligence (XAI). While there are already many designs, the popular approaches used for real-world deployments of XAI are expert-centric dashboards [1–8]. This belies a broad consensus that no single explanation is perfect [1], that there is no “silver bullet”.

This project, however, attempts to step back and consider, from a human-centred perspective, what actually comprises the important design elements for XAI that would be applicable to all users, covered by only a couple designs rather than a dashboard. To achieve this, the project focused on the following research goal: *Understanding and improving trust in computers through participatory design of XAI in a mobile banking app*. This provided an apt research question: *What are the critical design elements of XAI that promote acceptance and trust in a mobile banking app?*

1.1 Motivation

Currently, the field of XAI is still quite new and developing rapidly. There is, however, a notable absence of human-centred research in many of its areas, and its deployment remains guided by existing design principles without questioning their applicability for that context. While there will indeed be a great deal of horizontal transfer from wider user experience design and human-computer interface (HCI) work, we are concerned at how they are applied without confirming they are suitable, and without identifying possible exceptions to the accepted rules.

We are also highly concerned that this research and deployment is driven without a human-centred philosophy guiding the design process to ensure that trust is maintained and promoted between the user and the computer. Understanding and improving trust is extremely important for XAI, as one of the foundational pillars of XAI is the recognition that unexplainable AI are inherently untrustworthy, which is a major blocker on continued development and acceptance of AI research into new domains and applications.

At this moment, over 70% of people in Great Britain make use of online banking [9], with over a quarter of adults making use of digital-only banking such as Starling Bank or Monzo [10]. Given an increasing customer desire for improved service and convenience, banks are exploring computer automated decisions with XAI to meet this demand. Studying XAI within this lens is highly meaningful, as it covers a real-world, high-stakes, application of the technology. To this end, the project partnered with Starling Bank to explore these questions within the context of a mobile banking app.

Here, a participatory design process (co-design), would enable us to determine what was important for users in their acceptance of explanations given by XAI. As co-design is a human-centred approach to design, focusing on real user needs and desires, though would substantiate our findings as an early step in understanding the relationship of trust between users and computers as mediated by XAI, and the capabilities for XAI to promote trust in an honest and grounded manner.

1.2 Overview

In this dissertation, I record the process and findings of this research project into a human-centred approach to XAI in a mobile banking app. This **Introduction** (Ch. 1) is intended to outline the purpose and intentions of this project. Beyond it, I review relevant literature and discuss the **Background** (Ch. 2) of the study.

This dissertation then covers the **Methodology** (Ch. 3) used in each phase of the project. This includes reflection on how the studies performed seek to answer the research question, with consideration for the limitations involved.

This ultimately sets up the **Results, Designs, and Findings** (Ch. 4) that were made from thematic analysis of the studies, and evaluation of the XAI designs produced. This is rounded off with the **Conclusions** (Ch. 5) of the study, summarising what has been observed, what can be inferred, and what questions are raised for future research efforts.

Chapter 2

Background

To establish the relevant concepts and notions for this dissertation and the study described therein, this chapter discusses the background involved and a range of relevant literature.

Section 2.1.1 covers many of the notions that underpin the fundamental goal of promoting trust through developing human-centred XAI. While the studies performed have supported the implicit assumption that explanations improve trust between users and the system, they have more importantly revealed that this is nuanced and frequently dependent upon appropriate explanations for the intended audience.

In Section 2.1.2 we review some important developments in both XAI technique but also informative demonstrations of XAI, including potential concerns and limitations. Critically, these begin to identify fundamental design elements and principles, but also demonstrate that we need to consider explanation methods beyond seminal works such as SHAP [11] and LIME [12], as the nature of post-hoc analysis and explanation is found to run counter to the requirements for transparency for interpretable models in XAI.

Ultimately, we find that XAI research is undergoing a resurgent wave of developments that are increasingly human-centred. New work is critical of the assumptions posited at the outset of the field, and is beginning to develop more nuanced and grounded theories of XAI. This has been reinforced by renewed interest in “glass box” models, which offer transparency into their workings, providing internal explanations that are known-correct thanks to the transparency afforded; as opposed to obfuscated black box models, which require external explanations that attempt to demonstrate the workings. We are also seeing how human-centred, iterative design processes clearly address many of the issues faced in research, where designers are able to embrace the “surprising”

and “counter-intuitive” findings they discover. Overall, the outlook is hopeful and the direction for future research is positive, and we see that many of the concepts established and substantiated by this project are reflected by the wider research community, lending credence to our observations and suggesting replicability.

2.1 Relevant Literature

2.1.1 Trust in AI

A critical difference between much of the existing literature into user-machine or user-AI trust and what is relevant to this study is that we are not exploring users augmented with XAI. In a mobile banking app, users are provided with the conclusion, and hopefully explanation, for the bank / financial service’s decision process for a given scenario (i.e., loan applications). Hence, many of the studies into trust are not relevant, particularly when they measure participant trust from playing some game with or without XAI support (such as [13]). Similarly, many studies were directed at eminently expert-focused domains, such as developing machine learning systems [14], and so are not relevant in public-facing deployments. Such studies can still reveal important considerations, which have been included, however since producing statistically meaningful results would be harder without a gamified (X)AI user-augmented approach, it seems that there is currently an apparent dearth of relevant studies. With the onset of new regulations and the establishment and adoption of human-centred XAI and Fair Machine Learning guidelines [15–20] it seems reasonable to deduce that we will see an increase in relevant research and human-centred designs.

One of the most important studies in establishing how trust in AI can be improved is from Ashoori et al. [21], who performed a large scale study using Amazon Mechanical Turk [22]. Despite the typical concerns over studies using Amazon Mechanical Turk, the qualitative feedback gathered suggested that this study was comprehensive. Beyond establishing how explanations from transparent, interpretable models, particularly those that provide factual information such as “where the data came from”, was a significant component in promoting user trust, this study also provides an insight into the impact of stakes on the user’s interaction with and response to XAI. They demonstrate that higher-stakes scenarios lead users to consider “issues of empathy or morality”, which was not observed with lower-stakes. This correlates strongly with responses from participants

in our project, with similar conflicting opinions across a group, where a lack of human-like empathy, motivations, or understanding could either be felt as a positive or a negative, depending on the participant. Similar conflicts of opinions exist broadly on many areas shared with those observed in our project, especially regarding human error and bias (and the complementary machine error and baked in, automated, or systemic bias).

Another large scale study came from Yin et al. [23], which demonstrated that trust between user and system is especially dynamic for AI. They also conducted a large trial with Amazon Mechanical Turk, with just shy of 2000 completed entries, and showed a strong change in user trust after they had observed the “accuracy in practice”. This is especially pertinent, as there have been many demonstrated cases where AI performance in reality is significantly worse than in the lab setting [24]. This means that, in cases where an AI or machine learning system has some measure of uncertainty, then it is essential that this be communicated. It is possible that this is not included in the explanation, but instead used as part of an ongoing transparency and performance report that is easily accessible and well documented, so that users can observe the expected and actual accuracy over time.

Zhang et al. [25] found it was essential to provide measures of confidence with models, even when they used a local explanation method which plotted feature contribution towards the decision. It was discovered that participants primarily considered positive contributions, and not impacts of negative contributions to the local explanation. Whether this was indicative of an excessively detailed explanation, so was difficult to read, is hard to say, however it demonstrates that graphical explanations must be carefully considered before deployment. This study demonstrated that human-centred design of XAI is essential, as XAI designs may be accidentally misleading. Indeed, a more nuanced positive result, as explanations were found to increase trust, however there was a clear need for these to be well-designed and relevant to its intended users.

2.1.2 XAI Methods

Currently, the design, selection, and deployment of XAI is predicated by the understanding that no single explanation is suitable for all purposes, nor for all data. The existence of XAI books focusing on using a variety of explanations [26,27] supports this, as does research from IBM [1] and Microsoft [17].

This is reflected in surveys of the XAI software available, such as by Maksymiuk et al. [28], which also showed that XAI is becoming increasingly popular. This one in

2. Background

particular looked at 33 libraries, of which most popular libraries for Python had between 2k GitHub stars to over 12k. These also demonstrated that multiple explanations techniques are typically supplied and at least 17 of the libraries had multiple explanations methods, covering both local and global explanations, of which eight R libraries had at least 3 or more explanations, and five R libraries had at least four explanations.

Perhaps the most significant study in XAI design comes from Poursabzi-Sangdeh et al. [29]. They conducted a very large study with almost 4000 participants in total over four experiments with around 1000 participants each, requiring a very high approval rating for participation. The fundamental experimental design was to compare a glass box model to a black box model, and varied the number of features between them. Their results emphasised avoiding “information overload” and noted “the importance of testing over intuition when developing interpretable models”. An interesting design concept raised was to communicate to the user when data points were deemed “an outlier”, and withholding model predictions or explanations until after the user had spent time considering their own prediction. While this latter point is not directly relevant, we can consider it as evidence for making explanations potentially one step removed from the decision, however the difficulty in balancing this with the need to make it inline and thus easy for the user to engage with was not considered.

A seminal paper by Lipton [30] dives into the interpretability of XAI. One crucial finding Lipton made was how post-hoc explanations and analogy models could be extremely problematic. Lipton also recognised that a linear model is not always strictly more interpretable than a deep learning approach; we know now that deep learning models learn a kernel function that makes them equivalent to a linear model with complex kernel [31], a linear model with a highly pre-processed dataset [32], or even with only a single parameter [33, 34]. This strongly suggests to me that different methods for interpretability, particularly those with inherent algorithmic or direct local explanations should be explored, such as prototype-based approaches which can the exact data used and directly actionable steps for changing the outcome that tend not to be adversarial, unlike some image-based explanation techniques. This was expanded upon by Barredo Arrieta et al. [35] with a survey of common models according to its notion of transparency and need for post-hoc explanations or analysis. Barredi Arrueta et at. also explicitly explore the notion of the Target Audience as critical component of the XAI design, and establishes a comprehensive taxonomy based on an extremely detailed survey.

2. Background

Another seminal paper comes from Rudin [36], which provides astoundingly clear arguments as to why interpretable, glass box models should be prioritised over post-hoc explanation models. This also comes with recognition for how poor explanations can be highly misleading, and produce a “false sense of confidence” in a black box. While the focus is on high-stakes situations, this remains applicable for lower-stakes scenarios as well. An analogy for a central argument is that, since explanations for black boxes tend to have uncertainty (just as the black box they are explaining does) then they will essentially lie to the user uncontrollably and unpredictably. Since it cannot be fully understood from such an explanation, this means that the user is also unable to tell when the explanation is “lying” without relying upon different explanations (which simply reduce the likelihood without eliminating it). Rudin does explore issues with interpretable models and AI, and in particular finds that counterfactual approaches may be inappropriate for high-stakes decisions when used to explain black boxes. This is a particularly interesting point, as use of counterfactuals on interpretable models instead produces a net positive effect, especially when used with prototypical examples. This was observed as a key benefit with the ProtoPNet architecture [37], which was able to create an interpretable deep network that produces evidenced decisions or classifications using existing prototypes in the dataset.

Finally, Doshi-Velez et al. [38] produced one of the earliest papers that attempted to explore the nature of interpretability in AI comprehensively. While many of the concepts have since been expanded upon by the above literature and beyond, there are a few key notions that remain extremely relevant, yet are rarely explored. The most pertinent to this project was the inclusion of time-pressures as part of the XAI design process, described as “How long can the user afford to spend to understand the explanation?” This was very important for each of the XAI scenarios that were explored in this project, and has huge influence on what is appropriate for the XAI design. While human-centred evaluation can help determine what is a reasonable explanation, there must also be consideration for whether users need to understand the decision and explanation much quicker than anticipated, or in contrast when users are interested or invested and wish to “fully understand” the decision. Our studies revealed that users are intimately aware of these time-pressures and desires, and recognise where relevancy to, say, the stakes of the situation at hand and their existing financial literacy or technological expertise can hugely influence how long they are willing to deliberate.

Chapter 3

Methodology

This project aimed to utilise a participatory/co- design process to help reveal and substantiate critical design elements. To this end, a three-phase process was determined, where initial data-gathering focus groups (Sec. 3.1) would give way to ideation and analysis (Sec. 3.2), resulting in a series of designs (Sec. 3.3) based in or explicitly contrary to the focus groups. These designs could then be evaluated (Sec. 3.4), to elicit more in-depth and detailed thoughts and opinions from participants.

As there is little literature concerning human-centred design of XAI, it is appropriate to perform a more qualitative study in an attempt to surface what is relevant, rather than a more quantitative study that relies on existing theory and findings. While there has been an absence of specific research here, we do note that horizontal transfer from broader HCI plays a role in shaping the study and the methods used, however that these are generalisations implies that this study must also be ready to observe exceptions to the typical guidelines. This was essential, as producing XAI for important and high-stakes decisions must be done with an emphasis on correctness. Hence, this project aimed to produce substantiated designs relevant specifically to this field, here represented within the financial context of XAI in a mobile banking app, such that this could help guide and ground future work and implementations.

Here, we describe the methodology used at each stage and discuss limitations of this study and approach.

3.1 Focus Groups

To ensure a range of opinions, we organised two separate focus groups with different backgrounds. One was formed from members of the public (a total of 3), of which some were customers of Starling Bank, and all had online or mobile banking of some description, while the other was formed of Starling Bank employees (a total of 7), primarily fielded from the Customer Service operators.

This was organised as a semi-structured group interview with freeform sections, taking 2 hours (remunerated) for each focus group. These were divided into four main sections:

- General Experience with XAI: “What’s your experience with computer systems that make decisions?”
- Methods of XAI: “What are your thoughts on different ways of explaining decisions?”
- (A 10-minute break)
- Experience with XAI in a Financial Context: “What are your experiences and feelings on computer decisions in banking and finance?”
- XAI in Financial Scenarios: “Thoughts on some simple designs for some scenarios and interfaces”

Due to it still being 2021 as of writing, these focus groups were conducted via video conferencing. The full sessions were recorded (excluding breaks); with prior written and verbal consent. Following the sessions, transcripts were produced for use in analysis.

A slide deck was constructed to facilitate the layout of the sessions, containing only a few general questions and discussion topics for participants, with non-visible notes containing the remaining questions. Visual examples were provided or were available if appropriate, such as for common examples of computer-made decisions and explanations for decisions on common platforms, such as Facebook, Google, Netflix, or Amazon.

Each section was presented with minimal jargon and an emphasis on plain language where possible, such that the questions remained approachable to participants that had been explicitly advertised towards or selected from non-technical circles. The above sections were presented as their plain language form, however outside a few pre-prepared wordings, much of the structure was only intended as internally visible to the researcher, and so is presented below as such.

3.1.1 General Experience of XAI

This began as a roundtable of where did they think that computers made decisions, and what their experiences and opinions were. Following this, a few visual prompts were provided of real world scenarios, and the following questions were asked:

- Have you come across these? (Recommendations, ads, etc.)
- What’s your interaction with them?
- “Were you aware these are machine decisions?”
- Were you aware of any explanations for the decisions?
- How were explanations presented?

A range of topics and experiences came up, which were explored as appropriate. We had a considerable subsection regarding the “ideal” form of

3.1.2 Methods of XAI

This began as a roundtable for each participant to share whether they made use of existing explanations, and what their thoughts were on computer-provided explanations in general.

As this required a high-level approach to explaining different methods used for the decisions presented (and corresponding explanations), a more structured approach was taken regarding the questions. A goal here was having participants consider the “provenance” of a decision, and whether it factored into their opinions or not. We identified three key categories, with the further questions to explore the thoughts presented:

- Human-in-the-loop: “When these kinds of decisions are human driven? A playlist on Spotify or YouTube, curated by a person, perhaps?”
- Traditional Programming/AI: “When there’s a computer that has some checklist, going through a bunch of rules, and then trying to come up with the decision from that.”
- Modern AI / Deep Learning: “When there’s the feeling that the computer is doing something that isn’t a straightforward checklist, that there’s some kind of magic going on here.”

See Appendix A.1 for the full set of questions.

3.1.3 Experience with XAI in a Financial Context

This section began with a roundtable discussion of participants experiences with mobile and online banking and finance; initially, specifically as customers for the expert group, then opened to their full experiences.

While the general structure was broadly similar to the first section, more directed questions were asked, such as regarding whether there were higher stakes or not in this context. See Appendix A.2 for the full list. This included a few elements and questions from the Methods of XAI section, to help determine whether provenance of the decision and explanation potentially mattered more in possibly higher-stakes situations.

This included discussion of what could be considered the “ideal” experience. In particular, a goal of this was determining where it was felt that computer automation could be added, or augmented with XAI, that wasn’t already present. This also would indicate the levels of trust and the nature of that trust towards XAI, existing computer systems, and human-in-the-loop systems for both finance in general, but also within important or common tasks and user stories.

3.1.4 XAI in Financial Scenarios

For the final section of the focus group sessions, we looked at a series of more concrete examples of situations in which there may be computer-made decisions, or situations where there may be a human-in-the-loop with computer assistance:

- Overdraft/Credit Application
- Loan Application
- Password Reset
- Card Payment

These were determined in collaboration with Starling as potential avenues for using XAI and computer automation. They were presented in a randomised order for each session.

Each scenario was presented with mention to both a “positive” outcome (the card payment went through as expected) and a “negative” outcome (the card payment was declined). While there was an expectation towards disinterest when the system worked “as expected”, not interrupting the user from working on some other task that the

3. Methodology

banking/financial system is enabling (such as making a payment), there were potentially uncommon scenarios in which it could still be of value to the user, and it remains good science to account for the “blatantly obvious”.

A focus here was to determine what were the kinds of information that were desired, and general form of presentation, rather than A/B testing specific modalities or visualisations. This came from the desire to learn what was truly essential to users, a critical part of this study, rather than determine effectiveness, as user trust does not always account for what has been proven effective.

3.1.4.1 Overdraft/Credit Application

Applying for either an overdraft or for credit is, technically, regarded as a form of loan, however the facility was identified as substantially different in practice and serving different purposes from a loan. As these are also the predominant form offered by Starling, as it did not offer general loan facilities as of this study, it was used as a separate topic.

In addition, they have a “partial success” state unlike the other scenarios, in which you, say, received an overdraft but not to the full amount you applied for. As you could also be declined, this meant that there was a “positive” outcome that could clearly be of interest.

This scenario also opened the opportunity to consider with the focus groups what self-education could be provided that was relevant, such as over credit scores.

Participants were told that may consider the following information:

- Credit bureau data (i.e. credit scores)
- Income
- Purchasing & direct debit history
- Previous overdraft/credit

3.1.4.2 Loan Application

Loan Application as a scenario was developed as an extension to overdraft/credit with a critical difference, customised interest rates. As, for example, Starling’s overdrafts have only 3 specific interest rates that can be assigned as of this study, this meant that the degree of freedom was limited. With personal or business loans, however, the rate could have a much higher degree of freedom, and determining a more precise and accurate rate for

the customer could be highly beneficial. This was considered as both personal/business loans, though an expectation towards personal loans was emphasised.

This also opened the opportunity to directly consider the stakes for this scenario, as well as the overdraft/credit scenario, contrasting them. Doing so would serve to identify whether what kinds of explanation and their detail potentially changed depending on the stakes involved; in particular regarding the interest rate and the kinds of explanations deemed appropriate for it.

Participants were told that may consider the following information:

- Similar data to overdraft/credit
- That this is higher stakes
- Interest rates vs. income

3.1.4.3 Password Reset

This scenario was suggested by Starling, where users could record a video saying a supplied code-phrase, which would be then compared to existing videos and ID material/documents to determine whether it was indeed the user themselves. A model would also be present to confirm the code-phrase, however for the purposes of this study only the video portion was considered.

This was identified as a meaningful scenario as it represented a situation in which the user is quite likely to want immediate access to their account, and may be given stress over this, and so would want to unlock their account quickly. Currently, a process like is predominantly manual, with a human-in-the-loop verification. Since the model may make innocent mistakes, providing the user information to help them correct quickly this is ideal, or at least informs them of the issues prior to contacting a human for assistance.

3.1.4.4 Card Payment

Finally, this scenario was considered as a high-frequency set of computer-decisions and explanations. While already automated in many cases, it is unlikely that one is provided an explanation for when a payment is declined (except for insufficient funds in the account). For “non-trivial” situations, there are no currently explanations provided as to why a payment was flagged as, perhaps, fraudulent (and therefore blocked).

This scenario had the interesting situation of considering a “false positive” outcome, wherein fraudulent payments and transfers were not blocked. It was also particular for being a high-friction situation, as it was often desired for a payment to be made immediately, hence helping to identify the balance between security and convenience desired.

Participants were told that may consider the following information:

- Payment history
- Location
- Amount
- Recipient

3.2 Thematic Analysis

Once the focus groups were transcribed, thematic analysis began. This approach borrowed from the methods outlined as grounded theory [39] though was augmented with condensed notes to assist as I had no prior experience with such grounded theory.

In an atypical manner for grounded theory, I produced a semi-atomic set of codes that were hierarchically combined to form a bottom-up ad-hoc grammar. We then used these to identify a set of top-down axial codes that covered the major themes relevant to the research question revealed in the focus groups, though as it was an informal approach I did not devise a new set of open codes that reflected this.

While intended as a highly iterative process, the informal approach used above provided for a faster turn-around, however did not enable meaningful analysis of the codes themselves in a statistical sense. In particular, while there was useful interaction with the general context, and many ideas would surface between them, this also meant that there were many “atomic codes” that were not directly relevant to the axial codes, as they concern topics outside the conceivable context of banking/financial apps.

3.3 XAI Design

The design process was rapidly prototyped from there, with a few central ideas. First, produce a set of designs explicitly derived from the findings from the analysis. Second, produce a complementary set of designs that inverted one of the most important findings of

the analysis (the emphasis on textual explanations), such that we could probe this further. Third, ensure that these designs were based on appropriate methods of XAI, such that these could conceivably be used as part of a Wizard of Oz study. Finally, to apply these designs to a set of user stories that allows them to be understood in a more analogous and realistic manner, and also directly compared and contrasted by users in the evaluation session.

As part of producing believable prototypes, an attempt to replicate styling from the Starling Bank App would help reinforce the user stories and XAI choices. Due to unfamiliarity with graphical design, only certain surface level details were replicated, however assets from templates and screenshots of the Starling app were provided by and used with the permission of Starling Bank. These enabled a rapid turn-around and the ability to provide some contextual supplementary designs that reinforced these further.

This provided a total of four designs, split over two scenarios, Overdraft Application and Password Reset, and again across two core design elements, text and visualisation. Each scenario had a common pool of three user stories each, and so each user story was represented as both a textual XAI (inline with the findings) and a graphical XAI (contrary to the findings), to enable direct contrasting between the designs. The user stories for the Overdraft Application were called Alice, Benny, and Clarke, and the user stories for the Password Reset were called Dylan, Erica, and Faker (who was clearly an intruder).

3.3.1 Overdraft Application User Stories

- Alice is a Cybersecurity consultant wanting to expand on her photography hobby, so is looking to get an overdraft of £2000, however it only gave her £1800, so she wants to know what would be the easiest change to get her that last bit
- Benny is a university student wanting to extend his overdraft to £600 after a few too many nights out buying rounds, which he got but with a 35% effective rate
- Clarke is a manager at a local bookshop, and decided to try out after recently joining Monzo after recommendations from friends, and saw the offer for an overdraft on account creation; they were declined the overdraft since they are not looking to put much money into the account, and for very recently having a credit check

3.3.2 Password Reset User Stories

- Dylan looks completely different after lockdown, having lost weight and grown a bushy moustache
- Erica recently switched to contact lenses and has a cut on her chin from boxing, but more importantly has changed her hairstyle, dyed darker, with long bangs
- Faker is not the owner of the account and is trying to break in, but as it turns out has similar shaped ears (but not that similar) and very similar skin tone

In particular, this graphical vs. textual dichotomy was related to and could reinforce questions regarding provenance asked during the focus groups. XAI that produces text, here formatted as plain/natural language, typically needs to be “interpretable” in a classic sense of providing its own “chain of reasoning”. XAI that produces graphics, on the other hand, have wider coverage of XAI, including those that are not easily or exactly human interpretable, such as large deep learning networks. This also provided a chance to probe the dichotomy between local explanations and global explanations, in particular as potentially enabling different kinds of user agency and decision-making; for example, local explanations might more easily provide an actionable step the user may take, while global explanations may more easily reveal some unusual behaviour or bias in a model, which allows the user to escalate to a human and act as external verification for system fairness.

3.4 Evaluation

Once we had our four designs, the plan was to host a new pair of evaluation sessions for our original participants, or at least as many as could attend. Unfortunately, the expert group from Starling had a scheduling clash, and so only the public group session occurred, with a majority of its participants returning for a 1½ hour long session.

As I had two scenarios selected, the session was divided in half. First we discussed the designs for the Overdraft Application scenario, covering all user stories for each design, with the above questions asked either to prompt further discussion or to probe deeper. This was then followed by a discussion of the XAI techniques backing the visible explanations, presented in a high-level manner, to gauge whether this impacted their opinions. The second half was structured like the first, however covered the Password Reset scenario. This was presented as a slide deck.

3. Methodology

Each discussion was prefaced with a small contextual and visual introduction of how these occur, such as during account creation or during typical usage for the Overdraft scenario, or with a small example of the identification video recorded for use when a password has been forgotten or otherwise needs to be reset; and also used by Starling as identification when opening the account, which was how data is initially collected to match the user in the Password Reset video.

For each section, all user stories were presented prior to exploring the provenance of the explanation and the methods behind each design. As these were three designs to look at, participants were provided ample time to read each, and a freeform roundtable discussion of each design opened for semi-structured probing. This included questions related directly to the axial codes identified during analysis, which were then repeated in light of the methods:

- Appropriate Level of Detail in Explanation: Do you feel like this is an approachable explanation? Does it seem too concise? Would you like to see more detail? If so, what detail?
- Understanding in the Model: Do you feel you have a better understanding of how the computer made this decision? How it will make future decisions? Does that impact your opinion in it making a decision, do you feel you can trust it?
- Exploring the Explanation: Do you want to learn more? How would you expect to learn more? Would you like to drill deeper on one specific point? See more points and more of the decision process? What would you add?
- Actionable Explanation: Do you think this is actionable? What would you change? Compare and contrast to the other approach?
- Feeling of Agency/Control: What's your opinion on how much agency this provides you? What are the opportunities for control that you envisage from this? What would you add to get that feeling?
- Customer Education: What opportunities do you think there are? How might this serve that? What would you add?
- Managing Expectations: What is the expectation you have for this in terms of how fast and accurate it should be? What can we do to help people develop expectations that are in-line with the reality of these computer systems/decisions?

3.5 Limitations

This methodology had notable limitations, and future work would do well to mitigate and correct for these, and expand beyond them. As an initial study in XAI from a human-centred perspective, these limitations are primarily constraints of scale and scope.

Most notable was participant count, especially with the public focus group. Most notably is that this precludes meaningful statistical analysis of the work, leaving only qualitative findings. While we are not able to infer where this study's findings lie within the global distribution of public and expert opinions, we can still state that these findings remains valid and important. This follows from how the resultant effect within the context of ensuring correct and appropriate XAI usage outweighs statistical significance concerns, as these do reflect a subset of the experiences and opinions that will be encountered in reality, so they must be addressed regardless to ensure that coverage.

There also is the common concern over participant demographics. To collect members of the public focus group, first being university-wide circulation of the session and then circulation from the research group's social media was used. This implies a certain bias in the public participant group, however as we were using also experts from Starling and focusing on designs styled within that, hence leaned into a mobile-first culture, some of this bias remains inherent. One notable step taken was to advertise explicitly outside the computing departments. Future work should strive to mitigate this bias where possible, though as mobile banking is not yet ubiquitous there will remain some inherent bias. Similarly, experts from outside Customer Service, as well as experts outside Starling, would be desirable for future work, however this would be a significant endeavour.

Unfortunately, I was unfamiliar with thematic analysis, especially regarding grounded theory, prior to undertaking this study. This meant my analysis method was informal and suboptimal, and could not be used to derive any statistical meaning or perform clustering of specific codes or sections of the transcripts. As this technique requires experience and time investment for further iterations, future work would benefit from both of these.

The participant count issue reappeared, as we were unfortunately unable to organise for members of the expert focus group to return and evaluate the designs due to scheduling constraints. We were able to reconvene with the majority of the public participants, however some were still unable to attend. This means that the evaluation became more limited in scope, however the above point regarding the effect of the results still applies, and simply reinforces that more work in this area is required.

Chapter 4

Results, Designs, and Findings

4.1 Thematic Analysis of Focus Groups

As the research question of this project was to look at how we could understand and improve trust in AI, particularly through participatory design of XAI in banking apps, the thematic analysis of the focus group transcripts was central for determining exactly what the designs would be, as it determined what was deemed important. These themes (axial codes) were:

- Appropriate Level of Detail in Explanation
- Managing Expectations
- User Agency/Control
- Actionable Explanation
- Customer Education
- Exploring the Explanation

Of these, the Appropriate Level of Detail, Management of Expectations, and User Agency/Control themes were seemingly the most important, both by frequency and by how they came up almost universally from all participants. Providing Customer Education (generally about broader, relevant topics) and Actionable Explanations came up significantly less than these, however were still important themes as participants brought them up during discussion of the scenarios, particularly with respects to financial education and literacy. Exploring the Explanation came up the least, mostly as it seemed at

odds with many of the time-critical scenarios, or at odds with the presupposed Appropriate Detail, though it was notable for being included with higher-stakes decisions, and some participants were simply curious. There was also the situation in which the “problem” was the understanding of the task, externally to the model used, and so Customer Education with additional resources was prioritised over Exploring the Explanation.

Here we explore the analysis of the Focus Groups regarding each of these axial codes, expanding upon these and their impact on participant and user trust in XAI.

4.1.1 Appropriate Level of Detail in Explanation

Determining and presenting explanations with an appropriate level of detail was the most important of all themes. In particular, this was often expressed in terms of desiring a concise textual explanation in plain language. This was summed up by one participant as “Simple explanations, a couple bullet points.”

This has strong coupling with the other themes, which indicates that this is clearly of critical importance as it underpins many of the assumptions or desires for participants. Therefore, if one was to consider a “hierarchy of XAI needs”, this would form the foundation.

It was recognised by participants as a difficult balance to achieve, and where expanding on details as a form of simple exploration could be viable if it didn’t provide enough. This kind of expansion as a simple form of explanation was recognised by participants as a useful way to provide a concise form, but then expand with a few critical details for the interested. However, an observation that was made from repeated questioning is that many people are unaware of existing methods of explanation in the general context of XAI, as these are often hidden behind an additional interaction. Instead, based on participants’ experience with inline explanations such as from Netflix or Spotify, which were regarded as positive but overly concise, it seems that the apparent step would be to use a highly concise inline explanation whenever there is an apparent computer decision, which can then be expanded by a single interaction.

This was contrasted with long-form explanations, which were widely considered fatiguing by participants. An example of such fatiguing situations cited was Terms and Conditions, which participants admitted to skipping over. They specifically did not want this to happen, especially in situations that were felt to be important such as when making a Loan Application. Additionally, using a broad “#1 factor” for an explanation

4. *Results, Designs, and Findings*

was felt to be insufficient, with an example in the general context being given as music or video recommendations, which may claim a preference for a given artist or actor, or a specific genre, but which participants either disagreed with or simply wished to know what the actual songs or shows and films they watched which actually contributed to this. In particular, an example of providing a pairwise explanation which users were able to mentally interpolate between to determine whether it was relevant as recognised as useful. A worked example given was that rather than “You listened to an Abba song” which “tells me nothing”, it could instead give “You listened to Abba’s Voulez-Vous and Dancing Queen”, which was felt to provide an intersection where the participant said “oh yeah I see why it’s like those” and “it’s like this and that”, which seemed to resemble the subtler components of how humans give recommendations. This was important as one aspect in which computer decisions compared to human ones (or human-in-the-loop) was that there is an absence of emotion and context, and a sense of competency within the ability to understand the user. This comes more under the Managing Expectations theme, however providing that form of pairwise detail seemed to close this gap somewhat for the participant then brought it up.

A component of Appropriate Detail that is related with Managing Expectations was the participants’ requirement for transparency, naturally corresponding to how accurate and true the explanation is, and that it covers all actual components of consideration for the decision. In particular with recommendations, but also in general, both focus groups desired to know the “real reasons” behind a decision. In the general context, advertisements and similar recommendations came up as the most common example, likely due to the ubiquity of surveillance advertising when online, which was often described as “creepy” even when it was considered inaccurate, and how in many cases obscured why that advertisement or recommendation was made instead of something the participants felt was more relevant. Here in particular, the desire for transparency was desired, in knowing why that was chosen over anything else. This was also felt to extend to the financial context, where advertisements for financial services within, say, a mobile banking app, or recommendations for setting up “saving spaces” also warranted transparency as to why the user was seeing this in particular.

A component for transparency in financial scenarios raises an element of difficulty, in particular when considering the Password Reset scenario or with Card Payments, where “too much” transparency was felt to possibly expose private information and infringe

4. Results, Designs, and Findings

on privacy. For example, if the Password Reset scenario was unable to match you, it should not show a video of you as evidence for why it did not match, as that is leaking private (and clearly important) information that could help an attacker gain access to your account. There was also the component of transparency that was felt to be a concern for the bank, where it could expose their decision models and potentially allow people to game them, such as being able to avoid fraud detection systems. This was especially felt by the expert focus group, which paired with concerns of one expert participant that many advertisements could be scams thanks to a lack of provenance, which led to a general distrust for these recommendations.

Another component of transparency for the financial context was stating whether something was being “reviewed by a person or not”. This played directly into the subjective sense of trust, which was noted by participants in both groups to be something that changed over time, as they had become “more comfortable” with using mobile banking and considering the recommendations that they make.

In “high-friction” scenarios such as the Card Payment, where blocking the card on fraud protection ground causes an inconvenience to the customer by slowing them down, the key desire was for a single sentence that allowed the user to know it was for a “valid reason”. Where possible, these explanations should be unobtrusive and in-the-moment, allowing the user to identify whether the payment was actually made by them and why it was blocked. It was noted that, especially in the case of where fraudulent payments went through, users may wish to see an expanded explanation while in-app, after having frozen their card, to help with contacting the Fraud Protection services at the bank. This shows the context sensitivity in what the Appropriate Level of Detail is, and how there may be different explanations presented at different times to the user, based on where and when they are accessing them; i.e., an explanation brought up by a notification telling the user the payment was blocked, compared to going through the transaction history and expanding on a specific transaction.

In high-stakes and perhaps time-critical situations like the Password Reset scenario, where it could also be used by attackers in an attempt to gain access to your account, the ultimate balance for an explanation is “enough to help me, but nobody else”, as stated by a participant. Participants did not seem overly interested in the why, likely stemming from how this was seen as a highly complex task that would be difficult to explain, however the suggestion of simple “innocent” reasons for why it might be having difficulty recognising

the user were well-received, such as moving indoors to better lighting conditions, or removing a face mask. As part of transparency here, it was recognised that users may forget they had uploaded video ID, and so telling them that such videos were used may be required to “re-jog their memory”, which was hoped to promote the revelation of “oh, that’s how they know”. A recognised component was that the explanation cannot frame the decision negatively, described by one participant as “we know it’s not you”, as there might simply be such a radical transformation that even a human operator would double take, with one expert user describing such an event.

Less time-critical scenarios with significant stakes, such as the Overdraft/Credit Application, or higher-stakes (as indicated by participants) for Loan Applications, then we see a recurrence of the desire for bullet points, though participants emphasised the desire to see the information used by the system to make the decision. In particular with, say, a “partial success” Overdraft, where the user has been awarded it but with a slightly lower limit than applied for, or with a higher interest rate than expected, the user likely wants to know what were the deciding factors in why they did not get the full amount. With higher-stakes scenarios like Loan Applications, there was an emphasis on having “all the details”, as participants perceived this as an extremely deliberate and considered decision on the users part, and so wished for reciprocal detail from the system as to why they were, say, rejected. What was of note is that it’s possible users may not wish to learn the explanation if they were accepted out of fear of “upsetting” the system and losing what they got, so deciding they are “not going to risk it”, which is a much larger problem regarding policy that cannot be covered here.

4.1.2 Managing Expectations

The concept of Managing Expectations is multi-faceted and overlaps and interacts with many other themes, in particular with the Appropriate Level of Detail. This is notable, however, as it relates to many of the external observations that users and participants held prior to the explanation, and forms a critical component in the evolution of trust.

As a core example from the general context, which also influences the financial context, users recognised that there was a lack of transparency that belay a conflict of interest with many computer decisions. For example, social media such as Facebook, Instagram, or Twitter, and video platforms such as Netflix, YouTube, and TikTok, were felt by participants to prioritise engagement and advertisements over user relevant content

and posts. This led to a broad apathy towards such advertisements and recommendations among participants, and some noted that it even disrupted the user experience, such as needing to search for friends because they were not being seen in their social media feed due to promotion of celebrities. This seems to indicate that systems liable to preferential attachment were considered an overall negative for relevancy to the individual, and potentially indicative of conflicts of interest.

An important part of Managing Expectations also came from the observation that, perhaps due to explanations not being directly visible inline in many situations, participants seemed to rely on an internal justification for any decisions. This was most clearly demonstrated by the expectation that it “uses my browsing history”, and their history of interactions and engagement, and how participants clearly attempted to diagnose which of their recent actions, or what pattern of actions, were responsible for, say, seeing adverts for “dog training centres”. While some were proactive about this, and one participant claimed to regularly clear their browser history because of this, few were aware of how platforms would share information and how activities from friends on different sites could potentially influence their experience, or how fingerprinting means they can be tracked without browser cookies, history, or logins. This also seemed to be centralised on more discrete elements of interaction and engagement, such as videos or songs that were watched, whereas social media seemed more opaque and “creepy”, perhaps reflecting some awareness for how social networks are leveraged to extract more information than a browsing history could provide. A side effect of this is that some participants reported adverse changes due to one-off events, such as purchasing a gift for a family member, or looking up something due to a conversation. In these cases, users found it confusing that an exceptional interaction suddenly shifted the recommendations and such dramatically, despite it being contrary to the rest of their history, and noted that these seemed to last for far longer than expected. While this indicates a clear need for accessible User Agency/Control over such decisions, it also suggests that there is a mismatch between the impact of trends in history, which users seem to relate to, compared to a conflict in interest for promoting popular or high-margin products such as merchandise, which was reported by one participant. A combination of transparency and user controls, both fine-grained and broad-strokes, are clearly required. Noted was a similar situation in which music recommendations did not provide the ability to not see particular artists or genres, but was instead limited to specific songs, preventing the participant from

tailoring their recommendations to their actual tastes. While this is clearly a case of User Agency/Control, there is also the critical component that participants clearly expected these capabilities, and were confused at their absence. Therefore, there is a critical role in Managing Expectations here, as currently most large platforms are plagued by such gaps between their expectations and the developers' that leads to many support threads and community outcries which are met with stark silence or empty platitudes.

One unavoidable component of Managing Expectations for XAI is the perceived accuracy and bias that such as system "should" or "should not" have. As numerous stories of "biased AI" reach the headlines, this is a vital area that must be handled properly. This is an extremely delicate are of expectation, as participants displayed the full gamut from trusting in a computer to avoid human biases, or recognising that computers can end up with biases "baked in" without being able to "quite understand it". There were ambivalent participants, who considered that there would be bias either way, and focused instead on accuracy. Unfortunately, accuracy also has a spectrum between those concerned more over human error to those concerned over machine error. This was amplified by the desire for systems that were without "motivations", such as to ensure financial privacy and security, which contrasted against those desiring more "sentimental" decisions and understanding, be it in recognising context for financial decisions, or perhaps a friend's "wedding playlist" and other pieces of emotional attachment. Clearly, the only correct answer is the one for each specific individual at that moment in time, as balancing these would be highly subjective and potentially require cultural cues such as how private financial information is considered (such as public tax returns). Managing Expectations in this situation would likely require providing significant User Agency/Control, as only the individual can self-identify their feelings here, as attempting to be prescriptive could run foul of exactly the same problem as choosing only a single answer for everyone. User trust can increase over time, and their expectations can thusly change, and so it is important that they be able to update their decisions, however it also should not be forced or expected, as that could be interpreted as an attempt to lead them to a different conclusion, which again runs foul of the same problem.

A separate scenario as part of Managing Expectations comes from the time it takes for a decision. This is difficult to balance, as people have different expectations, and participants showed a range of opinions. Importantly, however, is that there is a clear trade between the perception of deliberation and consideration, and the speed of the decision. Many were

happy with the concept of instant “positive outcomes”, given that this is the convenience mobile banking is desired for. The contrast, however, was that people were concerned over how long the “negative outcomes” took. If a loan was rejected instantly, the expert participants felt that users might not believe that it had considered all relevant details. Rejecting a Password Reset video, or blocking fraudulent payments, on the other hand, need to be quick due to the time pressures involved. The overall conclusion was to err on security, and many were used to traditional banking decisions being measured in a number of days, so felt that taking 24 to 48 hours could be appropriate, even if the computer didn’t actually need that long. In cases such as Card Payments and Password Resets, erring on the side of security meant being faster, however, as it meant the user was able to contact Customer Services or Fraud Protection operators faster. If this emphasis on security is clearly communicated, then the user expectations, it felt, would be understanding towards any caution. There is a contrast in being overly cautious, which was noted as fatiguing, and so providing user convenience can still be desirable, however this is likely limited in capacity due to requirements for customer protection.

A final note on these kinds of expectations is that there was a wide expectation from participants that this more likely to be “objective” in a financial context, and that the capability of a computer to see “all” information and process it equally was recognised as a major boon here. This was then contrasted by an expert participant noting how sometimes the interface might not provide appropriate options or inputs that would enable this, or might not be able to extrapolate the future situation meaningfully given details such as pay rises due to overwhelming historical data. This is an area where transparency in what is used for processing, with explicit examples, would be critical to ensuring users have the correct expectations, and recognise that some situations like fraudulent payment not getting caught could be due to them not reporting a stolen card. For example, in the Loan Application scenario, one participant was stated that a computer determined interest rate may “cover a lot more factors than what a human could” and be able to state “this is what this customer’s interest rate should be for these reasons” with high accuracy, and “should not miss anything”.

Some expert participants expressed doubts that computers could handle tasks in their current workload, and so there will be expectations of what “only a human” can do, which will need to be addressed. Similarly, the inverse can be inferred, as there are tasks that they believe a computer could do, but in reality is unable, either for complexity reasons or

perhaps due to regulation. As this is likely informed by subjective experience, it could be difficult to manage, however it is an opportunity for Customer Education and transparency to help raise awareness for what is possible and what has been achieved.

4.1.3 User Agency/Control

The theme of User Agency/Control is an interesting one, as here we see a direct contrast between the general context and the financial context. In the general context, participants desired a range of granularities for their ability to control the decisions made. In the financial context, participants were primarily concerned with one form of User Agency/Control, the ability to decide whether a human or a computer was making the final decision. Usually, this surfaced as the desire to have a computer for the convenience of when it produces the “positive outcomes” quickly and securely, doing what the user intended, but then having the ability to contact a human in Customer Services that can override the computer when it is perceived to make a mistake.

This stands out unlike the other themes, as for the financial context of mobile banking apps, this seems to have a clear answer: provide a button to contact a human. There is more nuance, and other scenarios do have horizontal transfer from the general context, such as recommendations for services within the banking app, however all the posed scenarios were resolved for the participants through this manner.

An example of the nuance available is ensuring that users indeed have the full ability to enter relevant information, such as during a Loan Application. For example, the user needs to be able to select the actual reason for a loan, or to be able to describe it, however there are likely situations in which you either cannot select something appropriate, or there’s one that’s close but “not quite right”. When dealing with a human operator, these can be recognised, however a computer may struggle if there is a natural language option for explaining the purpose of the loan. While this is then resolved by indeed escalating to the aforementioned human operator, being aware of this within the XAI itself can be useful when providing the explanation and any feedback, such as prompting more appropriate categories for the reason.

An important consideration for User Agency/Control is that it must be convenient. Participants were in agreement that they did not approve of the dark patterns used to fatigue the user in the general context, such as when rejecting trackers and cookies, and felt that almost every situation raised in the general context should be Opt-In by

default, providing them control over their privacy. Some reported apathy towards this, as they believed it would “not stop the adverts”, or did not want to engage regardless and “feed the beast”. Since we are on the precipice of systemic change with ongoing investigations into many platforms, positioned for potential reforms and legislation, and a new wave of increased User Agency/Control through Opt-In/Out capabilities, means that these sentiments may soon change and users may feel empowered to exercise greater control; reports suggest 96% of iOS 14.5 users choose to opt-out of app tracking [40]. Such broad controls that do not duplicate effort, combined with existing fine-grained but effort-intensive controls, could change the public perception, as much of the existing tracking was described as “creepy” by both focus groups, and that one participant described going without their phone as “liberating”. Given that another participant decided to purchase YouTube Premium for an ad-free experience in an effort to “protect” their children, this is indicative that methods of User Agency/Control will be taken if they are believed and shown to be effective. With increased transparency, this may change computer decisions such that they are felt to be relevant to the user, with no conflicts of interest or sense that something is being taken. This could mean a reduction of apathy and increase in user empowerment, which participants suggested would promote greater trust. Concerns over privacy, which is User Agency/Control over their data and information, remains extremely important within the financial context, as personal information and financially sensitive data must be protected. Providing the ability to opt-out of services initially, whilst in the process of building a relationship of trust, then having the freedom to opt-in later when felt to be appropriate would fit accounts from participants in both focus groups.

4.1.4 Actionable Explanation

When it comes to Actionable Explanation, this theme has the greatest overlap with the Appropriate Level of Detail theme, however focuses on the ways in which a user is directly enabled to make changes that alter the decision. One participant described this as providing “clues” for what they need to do. This theme focuses mostly on the scenarios, perhaps appropriately considering how it is about identifying the concrete steps that a user can be directed towards.

For example, when it comes to the Card Payment scenario, the direct actions to verify a blocked payment or to agree to hold it were raised by the expert group, however the public group expanded on this with having “clues” to deal with situations such as moving

money into the account so that it has enough, or information on identifying fraud. This may be extended with direct access to freezing the card and reporting it as stolen, or contacting Fraud Protection or Customer Services.

An example raised for dealing with Password Reset videos is the forms of “innocent” correction a user can do: reminders to have good lighting, remove obstructions like a face mask or hair covering their eyes, or wiping down the camera. These were all noted as useful and actionable explanations that would allow the account holder to retry without needing to contact Customer Services and potentially have to wait in a queue. Given the possibly time-critical nature of this, it was commonly appreciated, however concerns were raised over the capability to deal with more significant bodily or facial changes. Given the concern over privacy and not wanting to provide potential attacks with usable information, it was suggested that this be balanced to provide limited feedback, and combined with this Appropriate Level of Detail be viable for XAI.

Another point to consider was raised for the Overdraft/Credit Application scenario, where direct action must be balanced with what’s achievable. In particular, it might not be reasonable to state a person should “get paid more” as this might not be viable in an appropriate timeframe. It was recognised that nuanced explanations may help, such as saying why a higher income would be important for that particular application considering the user’s other information. Ensuring that some steps are actionable may mean enhancing an explanation like “your credit score is just below the threshold” with “would you like some tips to immediately improve your score?”

Similarly, as was raised by participants in both groups regarding the Loan Application, users may be able to “do better”, perhaps given some information that was mistakenly omitted due to it not being considered relevant by the user. An example provided by a public participant was that of getting a loan with a higher interest rate “because you’re not a student, but I currently am a student”, and so the user can correct the mistake and improve the decision.

4.1.5 Customer Education

Opportunities for Customer Education refers to ways in which financial literacy, understanding of computer systems and XAI, and aspects of how Starling (or the particular banking/financial service or platform) operates. Only a few participants mentioned making use of additional resources and searching for common reasons why, say, overdraft

4. Results, Designs, and Findings

applications are rejected, and it is possible that there is some mutual exclusivity with how many participants prefer to attempt justifying the decision themselves rather than spent significant time trying to find an explanation. Therefore, it is important that such opportunities to provide Customer Education be made accessible, as it was noted by participants as one of the reasons for why their trust in mobile banking has improved over time.

In particular, regarding the XAI methods used, providing information on how the system works was noted by participants as likely helping inspire more confidence. It was, however, recognised as a difficult task in balancing between providing a meaningful explanation of the methods and in a way that was concise and avoided jargon. It is therefore more appropriate to consider this a goal of Customer Education, here with respects to “XAI literacy” or such, rather than a goal of the explanations given for XAI and computer decisions. Since concise text here is liable for jargon, opportunities for additional resources such as animations or videos may be explored.

A situation raised by the expert focus group was that many customers spend time waiting in a call queue for something that they could self-serve with the app, and that they simply do not know how. Thus, an element of “service literacy”, both with helping people realise what is available, but also in terms of how to use and navigate the app and services, would be an opportunity to empower users. A related idea was noted by the public focus group, where you could avoid queues in situations where the user “narrowly missed out”, such as on a Loan Application, which could provide financial advice relevant to finally getting the loan in a month or so, or to improve their interest rate when they reapply.

One clear example raised by both focus groups and widely desired was providing details on abstract information such as credit scores. While generally specific to the banking/financial service, many of the broader tips and financial advice applicable would improve general financial literacy. Some pieces, such as providing details on their profiling, were of interest to participants, but also met with a little hesitation from some, who felt it was invasive. This was also seen as an opportunity for transparency, such as reporting where information was coming from, and opening the door for improved awareness in where personal data has been collected.

4.1.6 Exploring the Explanation

The final theme, Exploring the Explanation, is notable in that it was seemingly less significant than the other themes, despite being a common component used in wider HCI fields that XAI draws upon such as in visualisation, where exploration is a popular tool [41].

One critical reason why this was not a more significant factor is the contrast with the User Agency/Control theme, which highlighted the ability to escalate to a human that can override a computer decision thanks to better understanding of a context or ability to consider information that the computer did not or could not. While some participants were willing to go to external resources and self-educate, most participants expected to be able to contact a human operator that could provide understandable and detailed explanations, where another human can determine the appropriate detail, and so they are able to explore the explanation in a conversational manner. This would suggest that it was felt XAI is unable to do such a conversational interface and explanation process, which was evidenced by participants dismissing chatbots as “going round in circles” and “drove me slightly bananas”. Hopefully future systems can improve on this, but currently this appears to be a common sentiment from participants. Without reaching for human-like conversational exploration, more typical methods of exploration can be considered, and one fundamental component of this in XAI is the availability of multiple explanation methods simultaneously. Since there is no “silver bullet”, it is likely that we may see a primary and secondary approach be recognised by users as sufficient, prior to escalating to a human.

One example of simple exploration being highly useful was brought up by the public focus group, where if the explanation is “You like crime dramas” then the user is able to ask “Okay, what have I watched?” and be provided with the relevant videos. This is then expanded by the opportunity to explore the contribution of a particular video, which then may provide a link back to it, or user controls to state their preference or the actual relevancy. In a financial scenario, this could resemble either service recommendations, or when it reports suspicious recent activity it shows the actual transactions considered suspect.

4.2 XAI Designs

To produce a series of designs for the Evaluation Session, I needed to transform the identified themes from the axial codes into some concrete design elements, and apply

them to some scenarios explored with the Focus Groups. Here, we selected the Overdraft Application rejection and the Password Reset video mismatch scenarios.

First, one of the most apparent results was that concise textual explanations, typically considered as a short list, such as bullet point, were a critical element. Alongside this, the text needed to be in natural language, given that the participants wanted a “simple sentence”.

Secondly, this choice is inverted, as if this is indeed critical to the design, then we should see a more positive outlook on the textual version, compared to a more graphical design.

Thirdly, we determined appropriate XAI methods that would inform the explicit design, such that they met the desired criteria.

Finally, each design was implemented over three user stories and provided with some styling reminiscent of the Starling Bank mobile app.

This meant we had two designs per scenario, which primarily contrasted in terms of the core design elements used and the XAI methods supporting them. These designs would be presented for each scenario as individual and mostly static, and so participant input could be gathered regarding desired User Agency/Control, and for Exploring the Explanation. Opportunities for Customer Education would be included where applicable, and participant Expectations would be considered in light of the XAI methods used.

4.2.0.1 Overdraft Application User Stories

- Alice is a Cybersecurity consultant wanting to expand on her photography hobby, so is looking to get an overdraft of £2000, however it only gave her £1800, so she wants to know what would be the easiest change to get her that last bit
- Benny is a university student wanting to extend his overdraft to £600 after a few too many nights out buying rounds, which he got but with a 35% effective rate
- Clarke is a manager at a local bookshop, and decided to try out after recently joining Monzo after recommendations from friends, and saw the offer for an overdraft on account creation; they were declined the overdraft since they are not looking to put much money into the account, and for very recently having a credit check

4.2.0.2 Password Reset User Stories

- Dylan looks completely different after lockdown, having lost weight and grown a bushy moustache
- Erica recently switched to contact lenses and has a cut on her chin from boxing, but more importantly has changed her hairstyle, dyed darker, with long bangs
- Faker is not the owner of the account and is trying to break in, but as it turns out has similar shaped ears (but not that similar) and very similar skin tone

4.2.1 Overdraft Scenario

4.2.1.1 Textual Design

This design (Fig. 4.1) uses a counterfactual approach to identify the necessary actions needed on the user’s part to change the outcome. This would make use of the user’s data, both what was input (such as their monthly income), and what can be collected from historical information and potentially external datum, such as recent credit checks against them. No ordering between counterfactuals is imposed.

The three user stories, Alice, Benny, and Clarke, are represented respectively in Figures 4.1a), 4.1b), and 4.1c).

We envisaged that a design like this would be supported by counterfactual local explanations similar to Fig. 4.2. One of the key benefits of this XAI is that it can produce a diverse set of complimenting actions that users can take to change the outcome of the model.

One limitation of this approach is that, with current methods, the feasibility of a change in reality is not considered. This means that if only a single set of counterfactuals is provided, then they might focus on the “simplest” change for the counterfactual XAI, but which is extremely difficult to accomplish for the user’s situation. For example, it might suggest an increase in income when the user has already recently received a pay rise. While these have been presented as if they are single-effect, it is possible that they may be based on multiple factors, perhaps ones expected to naturally compliment one another; for example, a decrease in spending alongside waiting 60 days. Multi-effect counterfactuals as part of the textual explanation, however, may increase complexity and harm conciseness, so were not used here as that would be at odds with the key design principle (concise lists of simple sentences), and are perhaps better explored interactively.

4. Results, Designs, and Findings

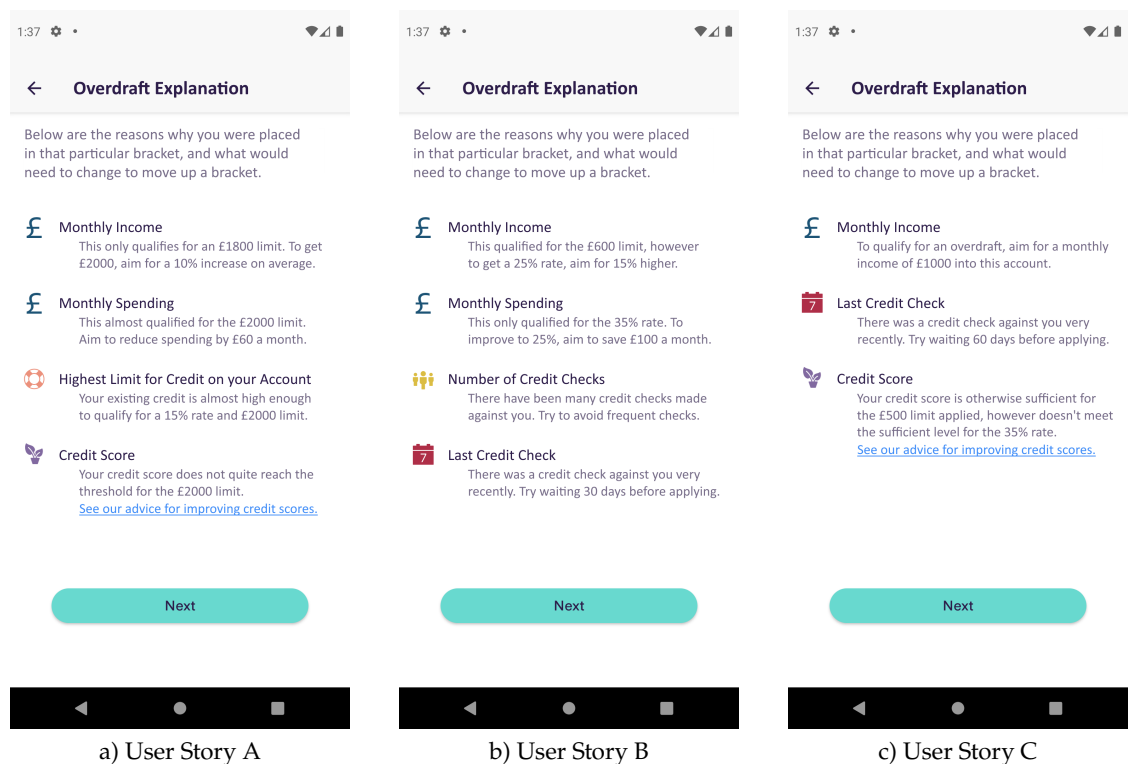


Figure 4.1: The Counterfactual Overdraft Explanation Design.

(b)	Original	CF
Workclass	Private	State-gov
Education	High school	Bachelors
Marital Status	Married	Married
Occupation	Blue-Collar	Blue-Collar
Relationship	Husband	Husband
Race	White	White
Sex	Male	Male
Country	United-States	United-States
Age	46	46
Capital Gain	0	0
Capital Loss	0	0
Hours p/w	40	40
Prediction	$\leq \$50k/y$	$> \$50k/y$

Figure 4.2: Example of Counterfactuals Guided by Prototypes [6]

4.2.1.2 Graphical Design

This design (Fig. 4.3) presents a series of graphical global explanations for each considered feature, such as partial dependence plots, which are zoomed and cropped to the locally relevant area: the assigned bracket and its immediate neighbours. Whilst the counterfactual approach presented only a set of “relevant” components, this is not inherent in this design, therefore we presented all available components such as their income, the number of credit checks against them, their monthly spending, and their highest existing credit limit. As the overdraft limit and the interest rate brackets might be decoupled, we envisage a touch interaction that switches between the two, with the default set to the most “relevant” of the two, namely where the user is closest to a neighbouring bracket. No ordering between graphs is imposed.

The three user stories, Alice, Benny, and Clarke, are represented respectively in Figures 4.3a), 4.3b), and 4.3c).

We envisaged that a design like this might be supported by an approach similar to Fig. 4.4, though we specify using PDPs explicitly, and despite similarity avoided including ICE to promote clarity given the small sizes. If only a single plot was displayed, this might have been changed, however the constraints of a mobile device’s screen place and the average reading distance a clear limitation on information density.

While we did not determine exactly whether to use a PDP or some other plot, the design element represent any graphical global explanations. This intended to promote greater awareness of where the outcome sat with respect to its nearby boundaries, which here would have been the current and neighbouring brackets for overdraft limit and interest rate.

A key limitation with this approach is that it requires literacy in this graph’s own design language, and so being able to recognise where user’s data was placed (here, within a circle), what the axis were, what the boundaries in terms of each bracket meant, and so on. Due to inexperience with producing visualisations meant for public consumption, the design was somewhat inadequate in communicating these, so it was explained during the Evaluation session. This resembles how such a visualisation might be viable when presented to the user by an expert in Customer Services, who can describe each aspect in a conversational manner.

4. Results, Designs, and Findings

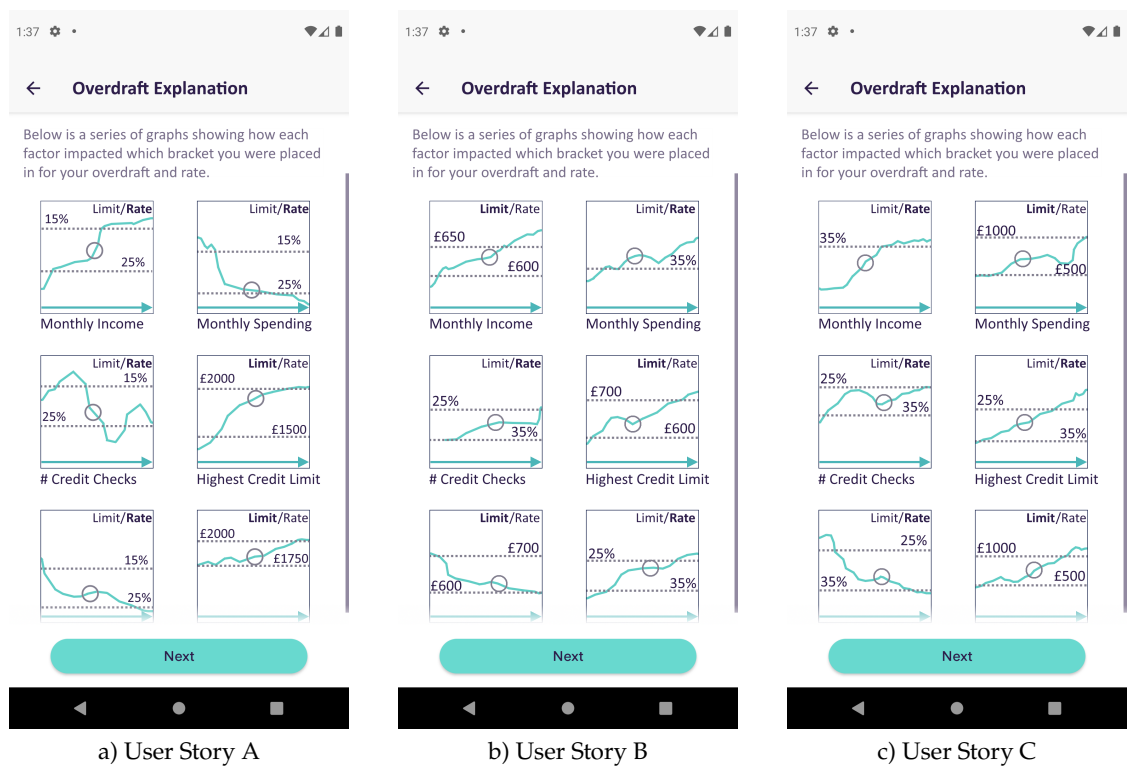


Figure 4.3: The PDP/Graphical Overdraft Explanation Design.

4. Results, Designs, and Findings

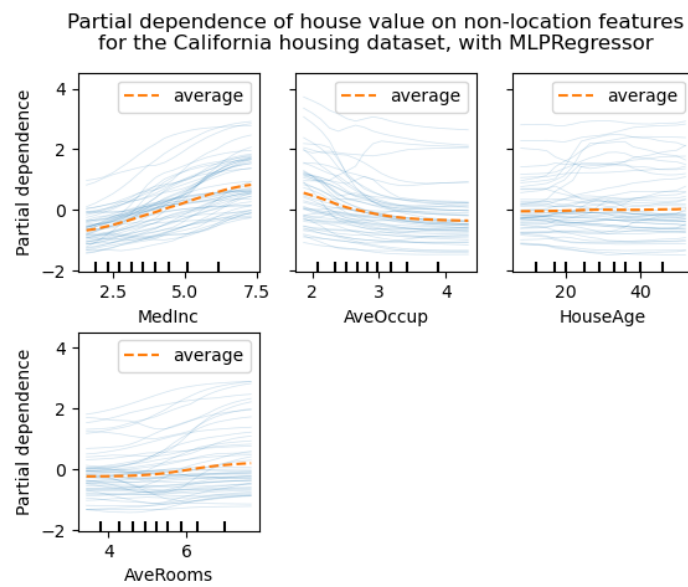


Figure 4.4: Example of Partial Dependence and Individual Condition Expectation Plots [42]

4.2.2 Password Reset Scenario

4.2.2.1 Textual Design

This design (Fig. 4.5) makes use of the binary state of local matching similarities directly to the portion of the user’s video dataset to highlight which components were non-matches, or combining potentially with heuristics to identify when they are obscured. This is based on an interpretable model, which is closer to glass box than black box with its presentation of important features. No ordering between matches is imposed.

The three user stories, Dylan, Erica, and Faker, are represented respectively in Figures 4.5a), 4.5b), and 4.5c).

We envisaged that a design like this would be supported by a prototype-based deep learning approach similar to Fig. 4.6.

Our design was inspired by Chen et al. [37] where they introduced a method for combining a prototype-based approach with deep learning to create an interpretable image recognition system. Since we would not present actual images, the idea was to use the matches, or absence thereof, to inform what features of the user’s face did indeed match the data already recorded for the account holder. This was presented as a binary match, though may internally use a confidence metric, however this was not presented as it would conflate with the graphical design for this scenario. This would mean that what the user saw were the features that either matched, or those that did not.

To ensure that this design only “recognises” the user, we would presume an XAI system that only can make prototyped-matches against the existing ID data available to Starling. This would include ID documentation, but also it would include the ID video made by the user on account creation, as well as any more recent videos they have submitted due to forgetting their password prior. This could potentially be used for broader security checks, and so there may also be videos from these. If the deep-learning component was originally trained to be able to perform matches on people agnostic of age and similar changes, then it may perform well even if the user only had their ID video from multiple years ago. If not, then it might not decide they are sufficiently similar, and would not find a match. This could be dealt with heuristically using style transfer approaches to synthetically age up older videos, or by recommending users upload new videos on an annual basis or so.

4. Results, Designs, and Findings

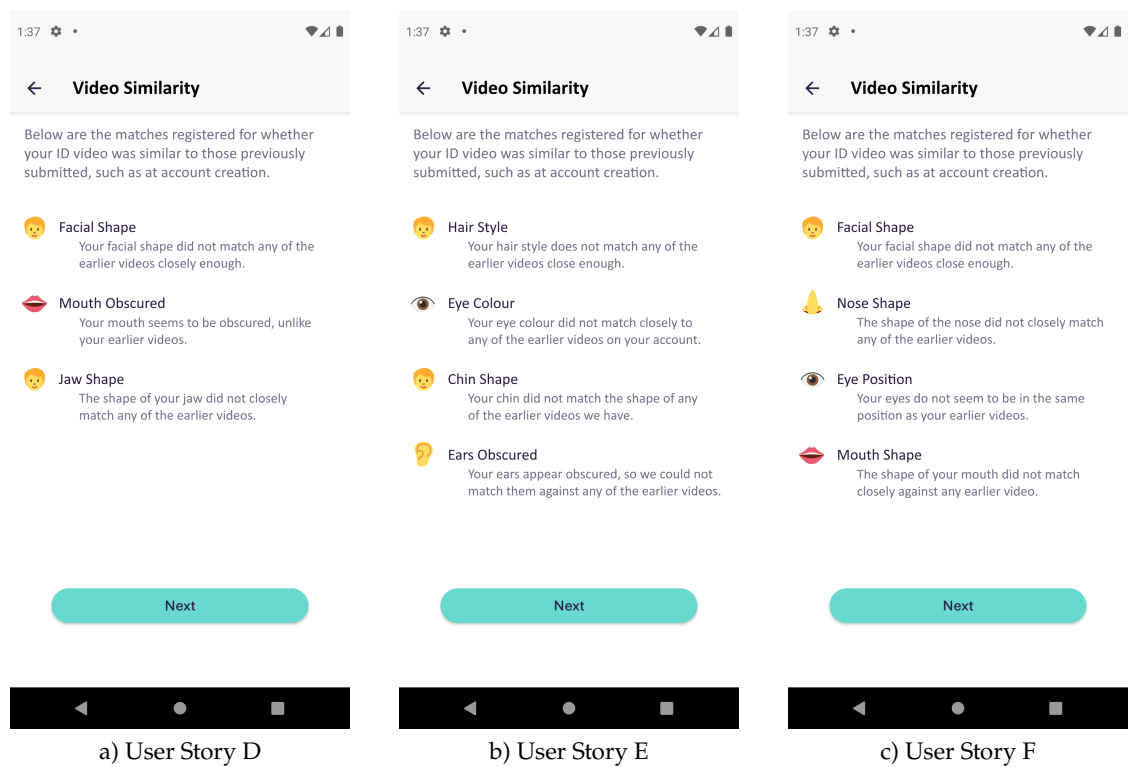


Figure 4.5: The Prototype Matching Password Reset Design.

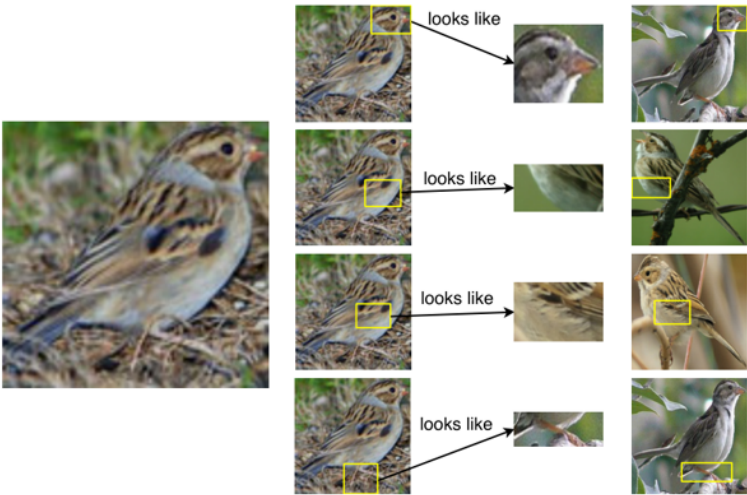


Figure 4.6: Example of A Deep Learning Approach to Interpretable Image Recognition [37]

4.2.2.2 Graphical Design

This design (Fig. 4.7) makes use of global 1D similarity scores for known features in an explainable black box design, typically a deep-learning approach such as labelled vector/encoding similarity or feature-aware discriminator. These are then ordered by importance to the decision, such that the first value was the most significant, perhaps seemingly regardless of its similarity. This ranking may be produced by a number of approaches, however single-effect strength is identified as likely the most meaningful within the context of appropriate and actionable feedback.

The three user stories, Dylan, Erica, and Faker, are represented respectively in Figures 4.7a), 4.7b), and 4.7c).

We envisaged that a design like this would be supported by a large deep learning approach similar at a high-level to Fig. 4.8.

We did not prescribe an exact method to be used, as this is meant to represent a more global explanation with some component of 1-dimensional confidence or similarity scoring. I combined this in a speed-dial visualisation with the global contribution concerned for determining overall facial recognition. This was colour coded and styled according to a figure used by Starling Bank.

The approach could be based on vector similarity, preferably with some known dimensions, or with a labelled encoding from an autoencoder. Alternatively, this could make use of a GAN discriminator, which could be trained in a one-shot manner on the user ID data and previous videos, and output a learned representation of their facial features. This might also use image saliency and attempt to determine global matching of features based on the previous data.

Since all features would contribute, a ranking was imposed based on importance to the decision. For the purposes of illustration, I roughly mapped this as using the highest confidence band for “Highly Similar” as most important, and then the highest distance from the peak of the band the “needle” is currently in to determine between similar rankings to represent nonlinear contribution to the final decision. In reality, this would likely be a more formal approach, such as permutation importance, or whichever global explanation fits the feature scoring best.

4. Results, Designs, and Findings

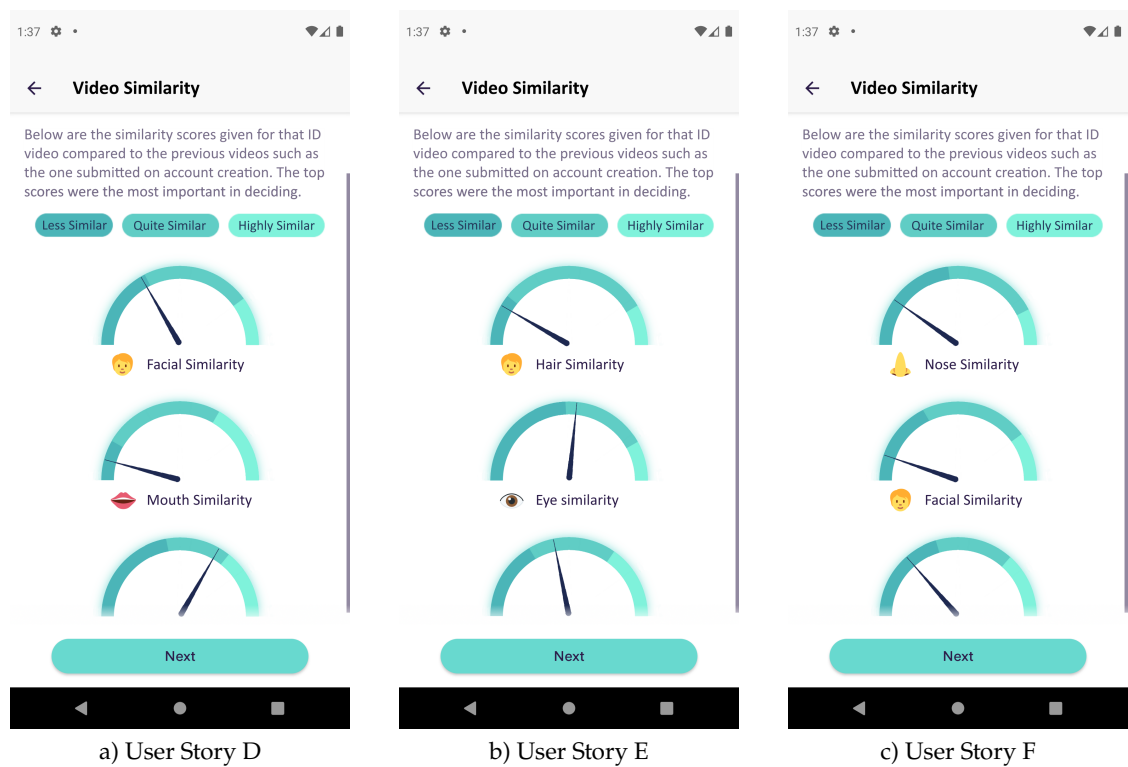


Figure 4.7: The Ranked Similarity Password Reset Design.

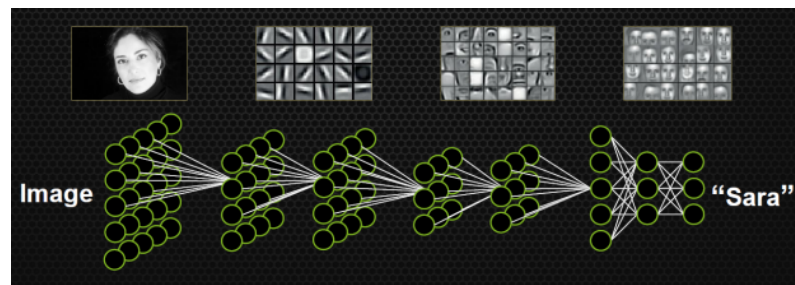


Figure 4.8: Example of A Deep Neural Network Approach to Facial Recognition [43]

4.3 Findings from Evaluation Session

The immediately apparent finding was the confirmation that the textual designs were generally more well-received than the graphical ones, corresponding to local explanations that were felt to be directly actionable. Participants could envisage graphical designs that could be useful, but overall they were considered supplementary to the textual designs. While this supports the expected findings, ultimately the more important components of this are understandings of why, and how can we refine our ideas of which design elements are important for XAI in a banking app.

This indicates that participants prefer interpretable glass box models, which are generally more amenable to such textual explanations, and by their nature are always able to provide an actionable local explanation.

4.3.1 Appropriate Level of Detail in Explanation

The textual explanations were recognised as meeting an Appropriate Level of Detail, which was supported by how participants, without prompt, engaged in an exercise of imagining what the next step was, and how the explanation could be used to form a plan within the banking app or for setting calendar reminders. This was very positive, as it suggested that a concise list of “bullet points”, consisting of a simple sentence or two each, would be a good balance for such explanations to default to.

The graphical design for the Overdraft Application scenario was “more overwhelming than the writing” for one of the participants. This supports the argument for concise lists of simple sentences, which on a mobile device screen would also be relatively sparse to aid readability.

There were felt to be appropriate visualisations for these scenarios, in particular the graphical design for the Password Reset was well-received, but even the Overdraft Application left participants imagining how a visual design could be brought up to support the written form. Notably, their improvements prioritised on the visualisation in terms of readability, changing to utilise colour to emphasise where user stories similar to Clarke or Benny could put the “extra necessary to qualify for a successful overdraft”. This suggests that plots are viable, however require good visualisation and “an explanation” so that users would know how to interpret them.

4.3.2 Managing Expectations

There were few findings that expanded or refined the concept of Managing Expectations, and the original theme was uncontested. We were able to reinforce our analysis when it came in discussing the amount of time that was felt to be the maximum acceptable for the scenarios. The participants decided they were happy with taking 24 hours for the overdraft to process, especially if it was seen by a person, if it was going to be rejected. The Password Reset videos were expected to be accepted or rejected practically instantaneously, which matched expectations.

An interesting observation is that there is a dichotomy between what is deemed hard by the user and what is deemed hard by the expert. This matches the common notion found in AI research and deployment, where the tasks considered simple by machines are those found difficult for humans, while those considered simple by humans are found to be extremely difficult for machines.

4.3.3 User Agency/Control

Consolidating the feedback gives rise to a triple dot menus or such in one corner, perhaps besides the page title. This idea was developed with participants to provide options for exporting the explanations to a local PDF, directly contact Customer Services, and accessibility options. Having convenient features such as automated password resets is nice, but there needs to be other methods for those that can't or won't use it, and they should still be convenient, it just might be less so.

The explanation export was a novel idea raised by a participant, who realised that this could be useful for referring to at a later date as part of a plan, or to reference when consulting with a financial advisor.

Being able to directly contact Customer Services from within the explanation, after perhaps waiting in a queue, was recommended as that way you could either get a conversational explanation, or directly escalate and potentially correct errors. While it was suggested that bias could be reported here, participants were not confident they would be able to recognise bias in such explanations, so even they could report it, they felt it was unlikely they would.

Accessibility followed from a participant recognising that colour-graded visualisation could be problematic, and that Starling has a very prominent style but might in-turn sacrifice readability. This expanded to include changing colour schemes, contrast, text,

and font size to be more readable. We can infer this includes text spacing and line spacing, as well as specific accessibility options for Dyslexia such as additional spacing after full stops and supportive fonts.

4.3.4 Actionable Explanation

This theme was refined by the evaluation session, and found to appropriately describe setting up “what comes next” for the user. Both “clues” for the user to begin to perform on their own, but also the possibility of assisting them with this directly by helping produce plans with a timeline or to set calendar reminders.

For the Overdraft Application, this was considered by participants as notable, but not in the expected sense. As expected, there was a desire for the wording to improve and provide clear directions, but more importantly there was a desire for a blend of User Control and Actionable Explanations in the form of creating plans and setting reminders. In other words, the level of information was considered sufficient to move on to “what comes next”, and while the Next button might move a user to such a page, it was also noted that the ability to create plans and set reminders could be part of the explanations directly, with a button alongside it. This is a particularly interesting finding, as it reinforces that this is indeed an appropriate level of detail, and suggests that providing an Actionable Explanation should also include at least some direct methods to begin that action, where relevant. This has expanded upon the concept of the theme significantly, and indicates that it could become a very practical and user-desired feature of XAI systems.

For the Password Reset scenario, this was focused more on what could be immediately doable for the user, such as brushing hair away or taking off sunglasses, or potentially moving to somewhere with better lighting and less noise. Beyond that, the most important situation was account recovery, and so the explanation was noted to serve to jog a user’s memory as to what may have changed about them recently, such that they know what they need to tell a human Customer Service operator. Both designs were felt to fit this need, though the concise clarity of the textual explanation was felt to be most sufficient.

4.3.5 Customer Education

It was made clear that Customer Education opportunities are highly dependent on the scenario at hand, and potentially the time-pressures involved. While the Overdraft Application was felt to be an appropriate opportunity, perhaps thanks to how it is

a deliberate and considered situation where the user would likely spend some time digesting the decision and any explanation, the Password Reset was not felt to be appropriate. Perhaps it was seen as outside the remit of a banking/financial service to expand upon, or simply that the participants were not particularly interested in learning how such a system could work. It may also be a combination with this being a fixed function of the overall system, whereas Customer Education with the Overdraft Application is about broader financial advice and literacy.

One participant was particularly interested in explanations providing direct links to more materials and resources, such as FAQs, webinars, blogs, and reports. This supports the Overdraft Application's textual explanation, which included such a link, and was included in that participant's thoughts for how to produce an improved graphical explanation. This participant did not respond positively to the idea that the Password Reset scenario would need to store past video for comparison, and felt it was infringing on their privacy and self-expression to be "expected" to have to resemble themselves from the past. Such concerns beg the question of viable privacy-preserving approaches and XAI, and what can be done to educate customers as to how their privacy is protected.

4.3.6 Exploring the Explanation

Exploration was noted as desirable for the Overdraft Application, where a user might "tap on monthly income, and then it would show you the graph", the other participant describing this as "all the information on one screen, and then you can click on that and open up a new screen with more visualised data", to expand and explore upon the textual explanation. This was not desired for the Password Reset, seemingly to be due to the time-critical nature of it, where neither participant was concerned with learning more of why, but would rather retry or escalate to a human in Customer Services to help resolve this.

This is consistent with the thematic analysis, and the capability for multiple explanation techniques or visualisations that are already used in XAI. It does, however, demonstrate that there is a nuance in actual implementation that means human-centred approaches may need to limit this to perhaps a primary and secondary explanation, and recognise when it is simply a distraction to the user, and should be avoided in favour of priming them to contact human support.

Chapter 5

Conclusions

This project has begun to tease out the important design components and notions needed for effective, trusted XAI, particularly in the context of mobile banking apps. By taking a human-centred approach and researching practical designs, this project shows that producing XAI for mobile banking that users can trust is feasible and achievable.

5.1 Contributions

We have conducted an early human-centred study into the design of XAI systems that users would trust for mobile banking. This explored different design elements, with designs that consider the dichotomy of textual vs. graphical explanations, local vs. global methods, and glass box vs. explainable black box. Here, we found textual local explanations from interpretable glass box models were practical, sufficient, and actionable.

We have determined that these concepts are likely to be important in this endeavour, and form a foundation for further research:

- Appropriate Level of Detail in Explanation
- Managing Expectations
- User Agency/Control
- Actionable Explanation
- Customer Education
- Exploring the Explanation

We have expanded on and refined these to produce a few possible rules of thumb that might be considered by those doing further design research:

1. Use a concise list / bullet points, one to two simple sentences each
2. Value transparency and security, communicate why something's done that way
3. Be concise and readable first, let it be explored/expanded later if needed
4. Always let the user contact human operators that can override the computer
5. Prioritise accessibility, make some convenience available to all
6. Trust users to make a plan from concrete details, help them achieve it
7. Some people want to know more, make it easy to find in formats they prefer
8. Users would rather talk with experts than fiddle, keep exploration simple

5.2 Future Work

This work raises many further questions for new research. Beyond refining this project with larger studies and more analysis, and testing different design elements with Wizard of Oz studies and evaluations, I believe there are important questions revealed for XAI at large.

Human-centred design of XAI is nascent, and as shown has important design considerations specific to the scenario it is deployed in. Considering broader design principles, potentially constructing design matrices and plotting where different XAI designs fall within them, would be of great value when trying to determine complementary XAI that could be used where one supplements the other, will likely transfer to many scenarios.

Finally, we consider the fundamental question of trust. In the end, trust is an amorphous and dynamic relationship, and if we wish to make use of computer automation to provide greater accessibility, convenience, and capabilities, then we need to truly consider what is necessary for trust in the population. This likely involves many facets and likely requires many changes.

Developing these concepts and researching into the actual human impact and concerns will clearly be critical for an increasingly digital society, as without such work we will simply have a fragile foundation upon which our architecture of trust is built. Best we iterate on these in research rather than endanger the finances of the public with risky decisions.

Bibliography

- [1] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, “One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques,” 2019. [Online]. Available: <https://arxiv.org/abs/1909.03012>
- [2] H. Nori, S. Jenkins, P. Koch, and R. Caruana, “InterpretML: A Unified Framework for Machine Learning Interpretability,” *arXiv preprint arXiv:1909.09223*, 2019.
- [3] P. Piątyśzek and H. Baniecki. (2021) Arena — Interactive XAI dashboard. [Online]. Available: <https://arena.drwhy.ai/>
- [4] h2oai. (2020) Machine Learning Interpretability (MLI). [Online]. Available: <https://github.com/h2oai/mli-resources>
- [5] O. Dijk. (2021) explainerdashboard. [Online]. Available: <https://github.com/oegedijk/explainerdashboard>
- [6] J. Klaise, A. V. Looveren, G. Vacanti, and A. Coca, “Alibi Explain: Algorithms for Explaining Machine Learning Models,” *Journal of Machine Learning Research*, vol. 22, no. 181, pp. 1–7, 2021. [Online]. Available: <http://jmlr.org/papers/v22/21-0017.html>
- [7] The Institute for Ethical Machine Learning. (2021) XAI — An eXplainability toolbox for machine learning. [Online]. Available: <https://github.com/EthicalML/xai>
- [8] explainX.ai. (2021) explainX: Explainable AI Framework for Data Scientists. [Online]. Available: <https://github.com/explainX/explainx>

- [9] Office for National Statistics. (2019) Internet banking, by age group, Great Britain, 2019. [Online]. Available: <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/adhocs/10822internetbankingbyagegroupgreatbritain2019>
- [10] C. Barton. (2021) Digital banking statistics 2021. [Online]. Available: <https://www.finder.com/uk/digital-banking-statistics>
- [11] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.
- [12] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. Association for Computing Machinery, 2016, pp. 1135–1144. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [13] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, "I can do better than your AI: expertise and explanations," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19. Association for Computing Machinery, 2019, pp. 240–251. [Online]. Available: <https://doi.org/10.1145/3301275.3302308>
- [14] J. Drozdal, J. Weisz, D. Wang, G. Dass, B. Yao, C. Zhao, M. Muller, L. Ju, and H. Su, "Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ser. IUI '20. Association for Computing Machinery, 2020, pp. 297–307. [Online]. Available: <https://doi.org/10.1145/3377325.3377501>
- [15] B. Shneiderman, "Human-Centred Artificial Intelligence: Reliable, Safe & Trustworthy," *International Journal of Human-Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020. [Online]. Available: <https://doi.org/10.1080/10447318.2020.1741118>
- [16] A. F. T. Winfield and M. Jirotko, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180085, 2018. [Online]. Available: <https://royalsocietypublishing.org/doi/full/10.1098/rsta.2018.0085>

- [17] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for Human-AI Interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. Association for Computing Machinery, 2019, pp. 1–13. [Online]. Available: <https://doi.org/10.1145/3290605.3300233>
- [18] L. Floridi, J. Cows, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018. [Online]. Available: <https://doi.org/10.1007/s11023-018-9482-5>
- [19] High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019) Ethics Guidelines For Trustworthy AI. [Online]. Available: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf
- [20] J. Fjeld and A. Nagy, "Principled Artificial Intelligence: Mapping Consensus In Ethical And Rights Based Approaches To Principles For AI," 2020. [Online]. Available: <https://cyber.harvard.edu/publication/2020/principled-ai>
- [21] M. Ashoori and J. D. Weisz, "In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes," *arXiv*, 2019. [Online]. Available: <http://arxiv.org/abs/1912.02675>
- [22] (2021) Amazon Mechanical Turk. [Online]. Available: <https://www.mturk.com>
- [23] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the Effect of Accuracy on Trust in Machine Learning Models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. Association for Computing Machinery, 2019, pp. 1–12. [Online]. Available: <https://doi.org/10.1145/3290605.3300509>
- [24] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A Human-Centred Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY,

- USA: Association for Computing Machinery, 2020, p. 1–12. [Online]. Available: <https://doi.org/10.1145/3313831.3376718>
- [25] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, “Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’20. Association for Computing Machinery, 2020, pp. 295–305. [Online]. Available: <https://doi.org/10.1145/3351095.3372852>
- [26] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. lulu.com, 2020. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/index.html>
- [27] P. Biecek and T. Burzykowski, *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. [Online]. Available: <https://pbiecek.github.io/ema/>
- [28] S. Maksymiuk, A. Gosiewska, and P. Biecek, “Landscape of R packages for explainable Artificial Intelligence,” *arXiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2009.13248>
- [29] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, “Manipulating and Measuring Model Interpretability,” *arXiv*, 2021. [Online]. Available: <http://arxiv.org/abs/1802.07810>
- [30] Z. C. Lipton, “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [31] P. Domingos, “Every model learned by gradient descent is approximately a kernel machine,” *arXiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2012.00152>
- [32] C. Paterson. (2021) We were promised Strong AI, but instead we got metadata analysis. [Online]. Available: <https://calpaterson.com/metadata.html>
- [33] S. T. Piantadosi, “One parameter is always enough,” *AIP Advances*, vol. 8, no. 9, 2018.
- [34] L. Boué, “Real numbers, data science and chaos: How to fit any dataset with a single parameter,” *arXiv*, 2021. [Online]. Available: <https://arxiv.org/abs/1904.12320>

- [35] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>
- [36] C. Rudin, "Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. [Online]. Available: <https://www.nature.com/articles/s42256-019-0048-x>
- [37] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This Looks Like That: Deep Learning for Interpretable Image Recognition," in *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf>
- [38] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv*, 2017. [Online]. Available: <http://arxiv.org/abs/1702.08608>
- [39] J. Lazar, J. H. Feng, and H. Hochheiser, *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.
- [40] E. Laziuk. (2021) iOS 14.5 Opt-in Rate — Daily Updates Since Launch. [Online]. Available: <https://www.flurry.com/blog/ios-14-5-opt-in-rate-att-restricted-app-tracking-transparency-worldwide-us-daily-latest-update/>
- [41] T. Munzner, *Visualization analysis and design*. CRC press, 2014.
- [42] scikit learn. (2021) Partial Dependence and Individual Conditional Expectation Plots. [Online]. Available: https://scikit-learn.org/stable/auto_examples/inspection/plot_partial_dependence.html
- [43] L. Brown. (2014) Accelerate Machine Learning with the cuDNN Deep Neural Network Library. [Online]. Available: <https://developer.nvidia.com/blog/accelerate-machine-learning-cudnn-deep-neural-network-library/>

Appendix A

Extended Questions for Focus Groups

A.1 Methods of XAI

- What if it was actually by a person? (Editor’s picks, curation, and such? Computer recommends, human decides?)
- What if a computer works through the problem like a person? (You want to find out your tax bracket? It should use the same checklist.)
- What if we don’t know how/why the computer did it, but can guess? (Side question, does “it’s AI” feel different from “it’s a computer”?)
- Would knowing this affect your response to the computer/decision?
- What if the decision had an element of randomness? Not for every system, but can be effective, like GPT uses it to avoid giving the same response every time to a question, and recommendations will want to find interesting new stuff at random since otherwise they might get buried.
- Have you been given explanations by a human? As a pop-up when looking at the ad or search result?
- What if it gives the ticked off checklist that it used to actually decide as a kind of report?
- (Or some relevant portion? Showing “oh, not this and not that, but instead these” where perhaps that’s calibrated by “global” importance, so you’re seeing “relevant”

- ones you didn't get ticked. Or "a list of the steps that the system went through to make the decision" which you can expand or collapse to explore, a bit like a doctor or nurse listening to you and saying "oh it might be this, or it might be that"?)
- What if we weren't given a report of the decision-making process itself, but instead were figuring out why a decision was made, like a review?
 - How much detail would you be interested in? It probably depends on the context and decision, so what situations might need certain amounts?
 - "To what extent and in what ways do you want over these kinds of decisions?"
 - They probably say "it depends." So what do you think is the appropriate amount for a given context. What should the interface be?
 - If it's a case like you're driving along and your car thinks your about to crash, should it be giving a lengthy explanation? No, if any then must be simple and straightforward, since there's not much time, and you need to react or brace.
 - But something like you have to decide to have an operation, you probably want a lot of detail and explanation, from human or computer, and you want control over the final decision.
 - "How would you like the explanations shown to you?"
 - Could an explanation be interactive, or should it be "matter of fact"? Where? In a notification?
 - Should decisions be made with a "chain of reasoning", or is it okay to just try and explain why it was made afterwards?

A.2 Experience with XAI in a Financial Context

- What are your experiences from mobile or online banking?
- What are the computer decisions you are aware of?
- What are the stakes in some of these decisions? Login for a branchless bank like Starling is obviously pretty important. Not having your card or details stolen.

A. Extended Questions for Focus Groups

- How were these decisions made? Were there explanations?
- Did a computer hand over to a human to make a decision or to provide explanation for you?
- How did it notify you? Was any explanation apparent or hidden?
- Would you prefer if there were explanations at all?
- Do you think that would change how you interact with or view it?
- Does the kind of explanation matter? Justification? In writing or more visual?
- What control/interaction would you like? Would that improve things?
- Should the decision have a "chain of reasoning", or would it be okay if it just explained itself afterwards? Does this depend on the stakes?