Bringing Transparency Design into Practice

Malin Eiband[1], Hanna Schneider[1], Mark Bilandzic[3],
Julian Fazekas-Con[2], Mareike Haug[2], Heinrich Hussmann[1]

[1] {malin.eiband, hanna.schneider, hussmann}@ifi.lmu.de
[2] {julian.fazekas-con, mareike.haug}@campus.lmu.de
[3]mark@freeletics.com

[1,2]LMU Munich, [3]Freeletics GmbH, Munich, Germany

ABSTRACT

Intelligent systems, which are on their way to becoming mainstream in everyday products, make recommendations and decisions for users based on complex computations. Researchers and policy makers increasingly raise concerns regarding the lack of transparency and comprehensibility of these computations from the user perspective. Our aim is to advance existing UI guidelines for more transparency in complex real-world design scenarios involving multiple stakeholders. To this end, we contribute a stage-based participatory process for designing transparent interfaces incorporating perspectives of users, designers, and providers, which we developed and validated with a commercial intelligent fitness coach. With our work, we hope to provide guidance to practitioners and to pave the way for a pragmatic approach to transparency in intelligent systems.

Author Keywords

Transparency; explanation interfaces; participatory design; mental models; intelligibility; scrutability; intelligent systems.

INTRODUCTION

Intelligent systems are becoming increasingly pervasive in everyday life: Voice-commanded virtual assistants, product recommenders, spam filtering applications, smart-home devices, and personalized news feeds are already state of the art and support users in various tasks. Intelligent systems track and process user and usage data, from which they learn and derive their decisions and predictions. How these decisions and predictions have come into place is often hidden from users, which has been shown to negatively impact user acceptance of system reasoning [7], and satisfaction with recommendations and predictions [15, 22, 25]. Moreover, trust in the system and its predictions is diminished in opaque systems [11, 29].

Over the last 30 years, researchers therefore have repeatedly called for more *transparency* in intelligent systems, and have presented design guidelines and exemplary prototypes for such explanations (e.g., [21, 27]). Yet, to date, these learnings have barely been taken up by industry and practitioners. A notable exception is the widely adopted use of labels in recommender systems [37]. Beyond that, the inner workings of most commercial intelligent systems, such as Facebook's news feed, the Google search, or the fitness app presented in this paper, remain opaque.

There are several possible reasons for the industry's lack of enthusiasm to make their applications transparent:

(1) Proposed guidelines are not easy to integrate into complex real-world scenarios since they partly remain on a very abstract level (as we will see later in this paper);
(2) they come with requirements conflicting with real-world conditions, such as the need for extensive screen estate to integrate explanations [21];
(3) work on how to integrate transparency into *existing* UIs is, to the best of our knowledge, still missing; and
(4) companies might use algorithms that, by nature, may only be made transparent to a certain extent, such as neural networks and deep learning.

Hence, industry might not have seen clear additional benefit that would justify the investment necessary to develop transparent interfaces. However, this situation will soon be changing tremendously: European Union's General Data Protection Regulation will become enforceable on 25 May 2018. It includes a "right to explanation" of algorithmic decisions [33] as well as a right to opt-out of such decision making altogether [34]. Furthermore, the need for transparency in intelligent systems has recently been expressed in the *Joint Statement on Algorithmic Transparency and Accountability* by the ACM U.S. Public Policy Council and the ACM Europe Policy Committee [5]. Hence, big and small companies will be faced with increasing pressure to make their systems transparent.

The aim of this paper is to provide structured support in this task, in which the complexity and challenges of real-world applications and the needs of different stakeholders have to be met: Users might want to understand the system's reasoning, but do not want to be overwhelmed by information. Companies might want to meet the regulations on transparency without unveiling the details of the underlying algorithm, and thus their intellectual property. Designers might be faced with constraints that come with corporate design guidelines, limited screen estate, or system-specific user flows, and might need to solve conflicting needs between users and companies.

IUI'18, March 7–11, 2018, Tokyo, Japan

ISBN 978-1-4503-4945-1/18/03…15.00

DOI: https://doi.org/10.1145/3172944.3172961

Hence, guidance in the form of *processes* which can be adapted to the setting at hand instead of abstract design guidelines might be a better support for designers. As the industry partner for this research confirms, there is a serious need for a structured process to develop intelligent interfaces to meet the new regulations. However, such processes have not yet been defined in prior work.

In this paper, we contribute a stage-based participatory process for integrating transparency into real-world applications, which focuses on the improvement of users' *mental models* when working with the system.The process has been developed and validated with the Freeletics Bodyweight Coach [12], a commercial intelligent fitness app that computes personalized workouts for its users. We developed it in an iterative and user-centered style in co-operation with the main stakeholders with the aim of practically applying the outcome of the process to the app. We suggest exemplary participatory methods along with each stage to facilitate the application of the process.

The process is divided into two parts. The first part defines the *content* of an explanation (*what to explain*), the second focuses on the *presentation format* of the explanation (*how to explain*). The following stages emerged during the six months we spent working on integrating transparency into the system:

(A) What to Explain: Expert Mental Model
The *key components* of the algorithm are summarized in a hypothetical, "optimal" version of a *user mental model*.

(B) What to Explain: User Mental Model
The users' current mental model is elicited, and any differences and matches with the *expert mental model* are recorded.

(C) What to Explain: Synthesis – Target Mental Model
Based on the *differences* between the *expert mental model* and the *user mental model*, users select the *key components* from the *expert mental model* that are most relevant to them in their preferred level of detail. These *key components* are added to the *target mental model* and determine the focus of the prototype design.

(D) How to Explain: Iterative Prototyping
Based on corporate design guidelines and the user flow of the system, possible visualizations of the *key components* in the *target mental model* are explored.

(E) How to Explain: Design Evaluation
Differences and matches between the *user mental model* and the *target mental model* are investigated in order to evaluate the prototype design.

With our work, we aim at making transparency design more applicable in real-world scenarios, and argue that a pragmatic view on transparency in intelligent systems and case-by-case learning are feasible and promising approaches towards the emergence of best practices.

BACKGROUND

In the following sections, we first introduce the specific problems at hand with the Freeletics Bodyweight Coach before presenting existing findings and approaches of related work with regard to transparency that might help tackle these problems by introducing transparency. As we will demonstrate, previous learnings are difficult to apply to the specific requirements brought by the Freeletics Bodyweight Coach, and thus partially lack practical guidance for the case at hand.

Introducing the Freeletics Bodyweight Coach

The Freeletics Bodyweight Coach is a personal fitness application that offers users personalized body exercising via their mobile phone. Users get new training plans on a weekly basis, consisting of different types of workouts. The training plans are the result of an AI calculation based on, for example, each individual user's profile (height, weight, BMI), goals (loose weight, gain muscle, etc.), preferences (training days per week) and fitness level.

However, the rationale of the AI is currently[1] hidden from users. This lack of transparency became problematic as some users imagined real human coaches assembling their workout and requested the "human" coaches to better adapt the workouts to their preferences through the customer service hotline. The expectations that come with imagining real coaches assembling workouts are likely not met and may possibly lead to a negative user experience. On the other hand, when the coaching algorithm actually selected shorter workouts for regeneration, some users amended the training regime and performed workouts for two or three days in one session – a potentially dangerous misuse not intended by the application that can lead to overtraining and injury.

In summary, the current UI of the Freeletics Bodyweight Coach does not optimally support the underlying AI concepts. The aim of this project was to integrate transparency in order to help users to better understand the exercising suggestions.

Improving Mental Models through Transparency

In scientific terms, the above mentioned flaws in comprehensibility are called erroneous *mental models* [6] of the fitness coach. Mental models emerged as a concept in psychology and cognitive science, and have since been widely used also in HCI [32]. When we interact with our environment, we build internal conceptualizations of the objects, systems or processes around us that allow us to explain and predict their workings [31]. This applies to less complex systems, for example, light switches, just as much as to more complex systems such as smartphones. Our beliefs and conceptions about how a system behaves and reacts guide our future interaction [32], that is, mental models evolve as users interact with the system. Individuals have different experiences and backgrounds, and thus also develop individual mental models of a system in the course of their interaction. However, most mental models are simplified representations of the actual system workings [32, 45] – which is sufficient if it allows users to comprise the majority of observed system behavior [30]. For example, users do not have to know how electric circuits work in order to successfully use a light switch. However, if mental models are erroneous, or do not adequately reflect the complexity of a system, users may experience difficulty in predicting and explaining the system behavior [31]. Mental models may therefore indicate usability problems.

[1] This work is based on the mid 2017 version of the Freeletics Bodyweight Coach, which is under continuous development.

Making an intelligent system and its underlying design decisions *transparent*, i.e., explaining how the system works, has been shown to improve users' mental models of that system [22, 23]. In opaque systems, in contrast, users are more likely to build flawed mental models [32]. Improved mental models contribute positively to user satisfaction and perceived control [22] as well as to overall trust in the system [29] and its decisions and recommendations [7, 40].

However, improving mental models comes with several challenges: Mental models are characterized, among others, by their persistence (they tend to be robust to change even if they conflict with actual system behavior) [30], incompleteness (they represent only part of the system functions in an abstracted way), and instability (users forget details about the system over time) [32].

Increasing Transparency with Explanation Interfaces

To increase system transparency and to allow users to build better mental models, a commonly used approach are so-called *explanation interfaces*. Explanation interfaces originated as decision-aids in expert systems [13], and have successfully supported users in building useful mental models in, among others, context-aware systems [28], recommenders [15, 35], and interactive machine learning [21, 22].

The usefulness of explanation interfaces depends highly on their design [15, 22]. Two main design decisions have to be made [15]: (1) "What exactly do we explain?" (content of an explanation), which we will refer to as *what to explain*, and (2) "In what manner?" (presentations format of an explanation), which we will refer to as *how to explain*. In our case study – and we suspect this to hold true for many other products – both questions are not easily answered. For example, the UI could indicate that age and gender, past workout times and a model of a training cycle (including high intensity and recovery workouts) are used to narrow down the set of workouts. Alternatively, it could indicate the machine learning algorithms and data sets that are used. Integrating all information in a high degree of detail would require a tremendous amount of screen space and likely overwhelm or annoy users who prefer a simple UI to get their main aim – performing a workout – done. To determine (1) *what to explain* and (2) *how to explain*, we therefore reviewed prior work for guidelines applicable to our case:

What to Explain?

A basic question when designing for transparency is whether to aim for complete transparency, or whether to select information that is most important or useful for users to understand. On the one hand, several scholars have argued for completeness [21, 24]. Kulesza et al. [21] found that completeness was positively correlated with improved mental models, and did not impair user experience or task load. They define more complete explanations as including more of Lim and Dey's [25] intelligibility types. These intelligibility types include possible questions that users might have about the system model (Why did the system behave in a certain way? Why did it not behave in an expected way? How does the system work? What would happen if a specific interaction took place? What else can the system do?), as well as the system certainty about a calculated decision, and inputs and outputs. Similarly, Kulesza et al. [21, 24] argue that explanations should be "as sound as practically possible", meaning that they should not give the impression of a simpler system logic than actually used. This claim is supported by Tullio et al.'s [45] observation that users are able to form a lay, but surprisingly accurate mental model of machine learning concepts. At the same time, Kulesza et al. [21] acknowledge that explanations should remain comprehensible and that soundness and completeness have to be balanced against a possibly overwhelming amount of information.

The latter point has been stressed by researchers who recommend not to aim for completeness, but to select relevant and important information. Herlocker et al. [15] found that explanations that are too complex might lead to decreased acceptance of a system. Lim and Dey [25] have argued that implementing all intelligibility types is "excessive and may even be detrimental". This view is shared by Schaffer et al. [40] who found that full explanation of a recommender's reasoning had a negative impact on user confidence and enjoyment, which is in line with early work on explanations [13]. Instead, they suggest to "carefully explain the search strategy of a recommender to users when this is appropriate". Dourish et al. [9] argue on the same lines that systems unveiling their underlying logic should do so by disclosing features deemed relevant and hiding unnecessary details. Bellotti and Edwards [1] call for appropriate abstraction from the system model when informing users about the underlying calculations. The necessity for abstraction has also been highlighted by Tullio et al. [45] who distinguish between low-level (e.g., information about system input) and high-level explanations (e.g., information about how system input is related). According to them, high-level explanations allow users to modify their prior beliefs about a system more easily, and are suitable to address the challenge of persistent mental models.

Conclusion: To this date, there is no agreement in prior work whether to include all details of system logic in explanation interfaces. Moreover, there might not be a universal answer to this question, but it might rather depend on the product domain, user groups, and the context of use. For example, Pu and Chen [35] found that the preferred level of detail of explanations in recommenders was dependent on the perceived risk associated with the product domain: Users were satisfied with a short explanation for recommended books or movies, but preferred a more detailed explanation for products like cars and houses. Similarly, Schaffer et al. [40] as well as Gregor and Benbasat [13] highlight the importance of considering the user group or individual differences between users when thinking about *what to explain*: Explanations turned out to be more useful when tailored to the interest of a specific user group, and may even reduce user satisfaction if not [7].

Reviewing guidelines and recommendations on *what to explain* did thus not reveal an agreed-upon answer, but rather highlighted the need to determine the necessary amount of information on a case-by-case basis. As this challenge will be faced by many designers of IUIs, there is a need for methods and processes that help designers to answer these questions for their specific case.

(2) How to explain?

Once it has been decided what to explain, designers need to integrate explanations into the existing UI and user flow. To the best of our knowledge, concrete guidelines for the presentation format of explanations are still sparse: Prior work differentiates roughly between text-based and multimedia [13] and suggests that "general rules for interface design are at present the best guide available for choice of presentation method". Pu and Chen [35] claim that text-based explanations should make use of conversational language. Lim and Dey recommend to "augment" explanations with visualizations [25]. Herlocker et al. [15] explored a variety of explanation types and presentation formats; their results suggest that simpler graphs are more appealing to users. Kouki et al. [20] investigated user preferences for visualizations of explanations in recommenders, and found that Venn diagrams performed best in comparison to other visual interfaces, but were restricted by the number of information items presented. However, their findings also suggest that plain text explanations may perform similarly well. Lim and Dey [26] presented a toolkit for automatically creating standardized explanations in context-aware systems. However, when the authors applied the toolkit in a later study, they concluded that it was difficult to translate the automatically generated explanations into textual descriptions [27].

Several scholars agree that providing explanations should be done in a way that reduces users' cognitive effort [13, 21]. From their literature review, Gregor and Benbasat [13] conclude that low cognitive effort is expected, for example, for case-specific rather than generic explanations. Kulesza et al. [21, 22] claim that in situ explanations during interaction are necessary to support the formation of useful mental models. They further assume that helping users might be most effective if explanations can be evoked on-demand and come in "concise, easily consumable 'bites' of information", supporting a continuous, iterative learning process.

Conclusion: Recommendations on *how to explain* currently remain on a rather abstract level and largely rely on designers to interpret and apply them to a specific case. To date, there is no consensus as to when to use text-based explanations or visualizations, and in which form. Again, the answer is likely dependent on the specific requirements of a particular scenario.

Open Questions

Designing explanations in intelligent systems is a challenging task, and there is already considerable and very valuable work which may be taken into account when designing for increased transparency and improved mental models. Yet, we experienced difficulties when we tried to apply prior learnings to the concrete case at hand since many open questions remain:

(1) As shown in the last sections, opinions and guidelines about the design of explanations and *what to explain* differ or even disagree in crucial aspects, such as the level of detail and abstraction of an explanation. Also, advice is often given in a very general way, leaving much room for interpretation with regard to a concrete case.

(2) There is little guidance yet with respect to the *presentation form* of explanations. There are currently no best practices for visualizations or text-based explanations. Moreover, prior work does not take into account real-world needs influencing *how to explain*, such as company-specific style guidelines or corporate identities.

(3) Although research has presented various prototypes for explanation interfaces (e.g., [21, 23, 27, 35]), there is no work on how to integrate transparency into *existing* applications yet. For example, one important question as part of *how to explain* would be *where exactly* to display explanations on the interface. Also, many prototypes and guidelines do not take into account interface restrictions found in real-world situations, such as limited screen estate (except for [35]).

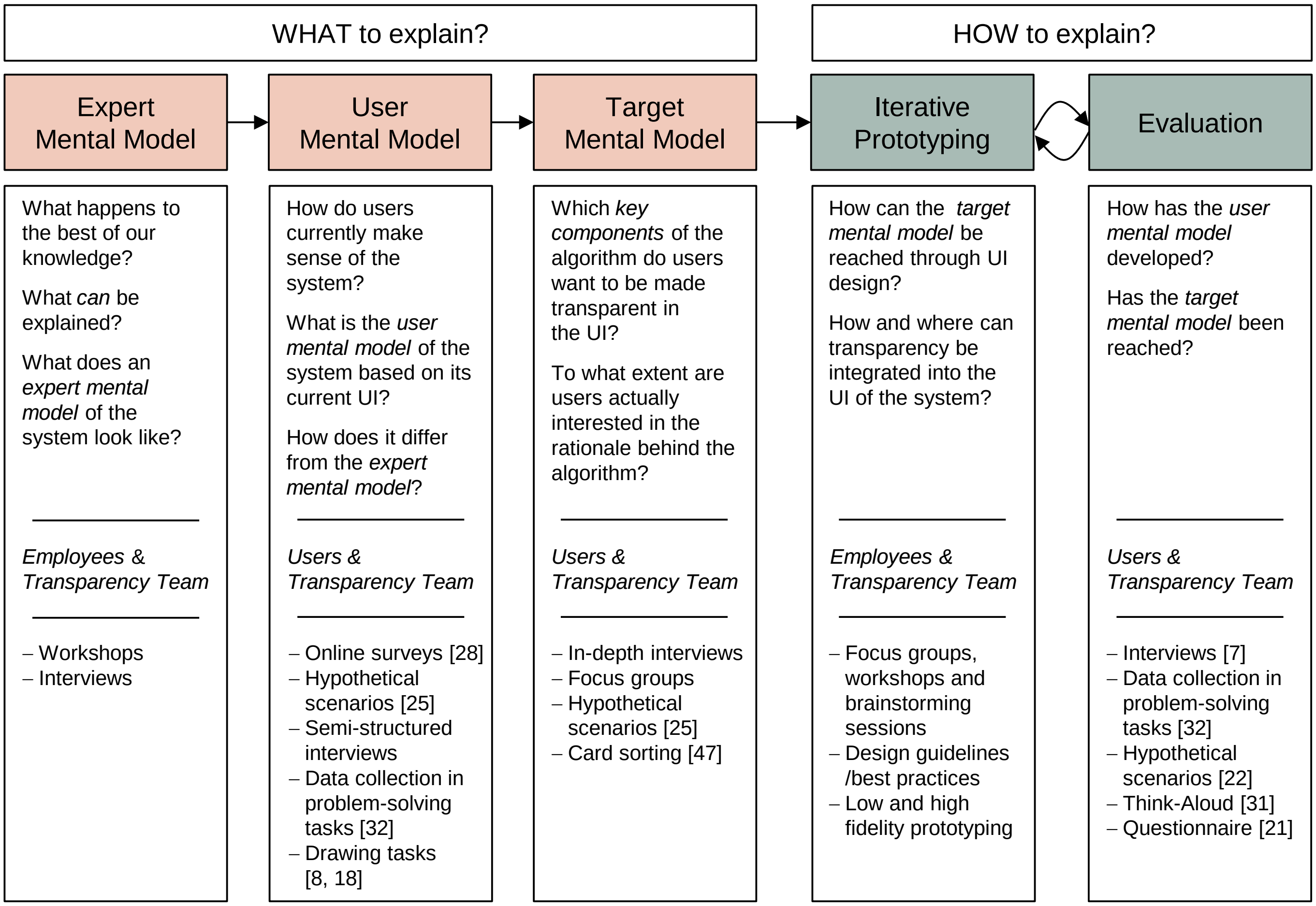
(4) In addition, designing for transparency comes with the challenge of meeting the different needs of the stakeholders involved in a real-world scenario. In our case, the project took place in a corporate environment, where design decisions in the product need to be aligned with several stakeholders in the company: management, marketing and branding, as well as the product design and engineering teams that consist of experts from sport science, psychology, computer science, machine learning, and UX/UI design.

We therefore argue that designing for transparency in complex real-world scenarios calls for individual solutions based on participatory design. Hence, the aim of this paper is not to formulate another set of guidelines adapted to a complex design scenario, but rather to suggest a *design process*, along with exemplary participatory methods. Designers can use this process as a guide in different scenarios that all come with different challenges to build appropriate explanations tailored to the target user group, while keeping the interface usable and in line with the existing interface elements and style guidelines. This is in agreement with calls for context-specific design solutions for explanations [13, 15, 35, 40] using a participatory approach [15, 25].

A STAGE-BASED PARTICIPATORY PROCESS FOR TRANSPARENCY DESIGN

In the following sections, we will present a stage-based participatory design process for integrating transparency into intelligent systems (see figure 1). The stages of this process are the result of six months work on a weekly basis in an eight-person team consisting of four external researchers from the university and company employees (two UX designers, one product manager, and one sports scientist). They are each guided by several central underlying questions that should be answered in the course of the stage. In line with prior work (e.g., [23]), mental models are a core aspect of our process.

The first three stages aim at clarifying the content of an explanation, that is, *what to explain*, the last two focus on the presentation, that is, *how to explain*. To answer the guiding questions in each stage, we suggest exemplary participatory methods that are established practice in participatory design, or are adopted from prior work on mental models. These suggestions are based on literature review and on the experiences made during our project, and are meant to facilitate the application of our process, but should not be taken as the only methods possible – other methods from the plethora of participatory design research methods [38, 46] might be equally appropriate.



WHAT to explain?			HOW to explain?	
Expert Mental Model	User Mental Model	Target Mental Model	Iterative Prototyping	Evaluation
What happens to the best of our knowledge? What *can* be explained? What does an *expert mental model* of the system look like?	How do users currently make sense of the system? What is the *user mental model* of the system based on its current UI? How does it differ from the *expert mental model*?	Which *key components* of the algorithm do users want to be made transparent in the UI? To what extent are users actually interested in the rationale behind the algorithm?	How can the *target mental model* be reached through UI design? How and where can transparency be integrated into the UI of the system?	How has the *user mental model* developed? Has the *target mental model* been reached?
Employees & Transparency Team	*Users & Transparency Team*	*Users & Transparency Team*	*Employees & Transparency Team*	*Users & Transparency Team*
– Workshops – Interviews	– Online surveys [28] – Hypothetical scenarios [25] – Semi-structured interviews – Data collection in problem-solving tasks [32] – Drawing tasks [8, 18]	– In-depth interviews – Focus groups – Hypothetical scenarios [25] – Card sorting [47]	– Focus groups, workshops and brainstorming sessions – Design guidelines /best practices – Low and high fidelity prototyping	– Interviews [7] – Data collection in problem-solving tasks [32] – Hypothetical scenarios [22] – Think-Aloud [31] – Questionnaire [21]

Figure 1. Our stage-based participatory design process for the integration of transparency in intelligent systems. The first three stages focus on *what to explain* in the system (content of an explanation) the last two on *how to explain* (presentation format). The stages are each guided by central underlying questions and involve different stakeholders. We also suggest exemplary methods for each stage that are either established in participatory design or have been used in prior work on eliciting and improving mental models.

Each stage may involve several stakeholders: members of the team responsible for the integration of transparency, other members of the company, and different user groups. We will distinguish between these stakeholders as *transparency team*, *employees* and *users* in the remainder of this section.

We will furthermore refer to central aspects of the algorithm, be it input items, output items, the relation between those items, or calculation steps, as *key components* of the algorithm.

Complementary material as to the application of the process in the Freeletics project can be found under the following link: medien.ifi.lmu.de/team/malin.eiband/transparencydesign.

What to Explain: (A) Expert Mental Model

The first stage serves two purposes: (1) The *transparency team* acquires knowledge about the system logic through communication and exchange with *employees*. (2) From this knowledge, the *transparency team* extracts the *key components* used in the calculation of the algorithm to build what we call an *expert mental model*, a hypothetical version of a user mental model that includes all *key components*. This is likely to require a certain level of abstraction from the system logic, and may take into account intellectual property protection.

Guiding Questions

What happens to the best of our knowledge? What *can* be explained? What does an *expert mental model* of the system look like?

Why is this Important?

The *expert mental model* serves as a reference for eliciting users' mental models in the next stages.

Outcome

The outcome of this stage should be twofold: (1) A *shared understanding* of the data collection and processing methods in place among all members of the *transparency team*, as well as a common language when talking about the algorithms. (2) An *expert mental model* that specifies all *key components* used by the algorithm, as far as possible.

Exemplary Methods

- Workshops with *employees* (approach taken)
- Interviews with *employees*

Application to our Case

To gain knowledge about the algorithm of the Freeletics Bodyweight Coach, our transparency team was invited to the Freeletics facilities for a workshop with the AI developers and other Freeletics employees. We were first given a presentation about the key components of the system logic and the wording used in the company to talk about them, and could then discuss any questions with the present staff members. We wrote down the key components, but we were also given a summary of the presentation and complementary material about the algorithm after the workshop.

From these insights, we created an expert mental model which we discussed in the transparency team in another meeting. The resulting version of the expert mental model was then sent to the developers for corrections.

The final expert mental model specifies five types of data that the AI uses to personalize users' training plans, namely athlete profiles (e.g., gender, BMI), user feedback (e.g., self-assessment of the workout technique), fitness level (e.g., results of an initial fitness test, performance class), training cycle (e.g., muscle group, day and week within the training cycle), and situational variables (physical limitations, equipment). Out of these key components, the transparency team identified (a) performance class and (b) training cycle as components with the strongest impact on personalization of training plans: (a) Every athlete is assigned to one of nine performance classes by comparing their workout times with workout times of athletes with similar profiles. Better workout times lead to a higher performance class and in turn to more difficult workouts in the workout selection process. (b) Every athlete's training plan is based on a research-informed twelve weeks training cycle. The training cycle is designed to improve the overall training results by alternating the training intensity and volume and incorporating breaks for regeneration. Based on insights in sports science, it is assumed that training success depends on the athlete following workouts and breaks as suggested.

What to Explain: (B) User Mental Model

The second stage focuses on user research and consists of two steps: (1) The *transparency team* elicits the mental models of *users*, based on the current UI of the system, and condenses them into one overarching *user mental model* representing the mental model of the target user group. (2) The *transparency team* then compares the *user mental model* to the *expert mental model*, and identifies any differences and matches with regard to the *key components* (missing, erroneous and correct *key components*).

Guiding Questions

How do users currently make sense of the system? What is the *user mental model* of the system based on its current UI? How does it differ from the *expert mental model*?

Why is this Important?

Users' current beliefs about the system logic are used to indicate the status quo of the current system transparency. This might depend on the target group and vary within different user groups of a system [13, 40].

Outcome

The outcome of this stage is again twofold: (1) The mental model *users* currently have of the system's workings, that is, the items that users identify as *key components* of the algorithm. (2) A list of differences and matches in comparison to the *expert mental model*.

Exemplary Methods

- Online surveys [28]
- Hypothetical scenarios [25]
- Semi-structured interviews (approach taken)
- Data collection in problem-solving tasks [32]
- Drawing tasks [8, 18] (approach taken)

Application to our Case

We conducted semi-structured interviews to investigate users' current mental models of the Freeletics app. Questions focused on (1) users' awareness of different key components of the AI, and (2) their idea of how these key components are used to generate their personal workout. The interview also included a drawing task, which has been used by other scholars to elicit mental models about privacy [18] and the web [8]. We also collected general feedback on app usage. All interviews were audio-recorded. These methods were selected because they allowed us to investigate users' mental models as they emerge through long-term product use while the algorithm processes their *own* data as input (in contrast to hypothetical scenarios [25]) and to generate in-depth qualitative data (in contrast to online surveys [28]). The interview was conducted with 14 active Freeletics users, ten men and four women between 20 and 42 (*M*=30.3 years), whom we recruited in a city park which is known to be a popular workout spot.

We used Thematic Analysis [3] to analyze the collected data. After transcribing the audio files of the interviews, we reviewed the data, searched for patterns, and then clustered answers and statements. From our data, we extracted the key components of users' current mental model of the Freeletics Bodyweight Coach. We found that the mental model already correctly included a part of the algorithmic key components, namely those that were explicitly shown in the current UI. For example, most participants suspected that their age and gender as well as past workout times might influence the training plan somehow. However, they had no understanding of the connections and relations between these components. For example, all participants were unaware that workout times of the last six months were used to calculate the performance class, which was then in turn used to calculate their training plan. Understanding this mechanism might show users that they need to improve their workout times consistently over a longer time to receive more challenging workouts in the future and that an unusually slow workout would not immediately impact their training plan. Hence, it might have an impact on users' expectations, patience, persistence, and ultimately, the user experience. Moreover, participants were unaware that their training plans follow a twelve weeks training cycle. Low intensity workouts (intended to incorporate recovery times) led some users to question the quality of the training plan and thus to train more than suggested. Revealing underlying principles from sports science in the UI might therefore convince users to adhere to the suggested program and to incorporate appropriate recovery times.

What to Explain: (C) Synthesis – Target Mental Model

In this stage, (1) users' explicit opinion of and interest in the *key components* of the AI is investigated and (2) the *transparency team* finalizes *what to explain* in the system logic: First, the *transparency team* uses the missing or erroneous *key components* (in comparison to the *expert mental model*) in the *user mental model* as a basis to investigate whether and in what degree of detail users are actually interested in knowing and understanding these components. When investigating users' explicit opinions, the *transparency team* can either use a within-group design (involving the same *users* as in stage B), or a between-groups design (involving different but comparable *users* than in stage B). The *key components* identified as important and valuable by *users* or the *transparency team* are then combined with the already correct *key components* in the *user mental model* to form the *target mental model*. This means that users either explicitly expressed their interest in this information or the *transparency team* has reason to believe that incorporating the information would improve the user experience.

The *target mental model* specifies which *key components* of the AI to focus on in the prototype design.

Guiding Questions

Which *key components* of the algorithm do users want to be made transparent in the UI? To what extent are users actually interested in the rationale behind the algorithm?

Why is this Important?

Since screen estate is limited, the *transparency team* is faced with the trade-off between displaying more information (higher transparency) and visual clutter or cognitive load for *users*. Moreover, it is likely that not all information about the system logic is perceived as relevant or helpful by *users*. While stage A and B provide useful insights into what can be made transparent, we argue that it is also important to elicit and acknowledge users' explicit opinion and interest in *key components* before jumping into the redesign of the UI design.

Outcome

The outcome of this stage is a *target mental model*. The *key components* from the *target mental model* that are still missing or erroneous in the current *user mental model* will be implemented in the prototype.

Exemplary Methods

- In-depth interviews
- Focus groups
- Hypothetical scenarios [25]
- Card sorting [47] (approach taken)

Application to our Case

We decided to let participants do card sorting [47] to find the key components most relevant to them. In the transparency team, we prepared 19 cards with text statements about key components in the target mental model in two levels of detail and conversational language, thus following Tullio et al.'s [45] and Pu and Chen's [35] recommendation. Cards with low level of detail only mentioned *that* a specific key component has an impact on the selection of workouts. Cards with higher level of detail explained *how* a specific key component influences the selection of workouts. Text-based explanations on cards allowed us to illustrate how each of the key components could be explained while avoiding that specifics of their visualization and implementation in a prototype influence users' reactions.

We recruited eleven participants in total, four women and seven men between 24 and 60 (M=33.5 years). They were a mixture of active long- and short-term Freeletics users, former Freeletics users, and people who had not used Freeletics before. Because we could not recruit the same set of users as in stage B, we incorporated some questions to elicit participants' current mental model (from stage B). Participants were then introduced to explanations on cards and instructed to sort them with respect to what would interest them most and help them to understand why a specific workout was recommended by the app while thinking aloud. They were also allowed to leave cards aside if the information seemed unnecessary or irrelevant. For each participant, we took a photo of the final card sorting and audio-recorded their verbalized thoughts.

For the analysis, cards were categorized as high-rated, medium-rated, and low-rated depending on how often they appeared in the participants' sorting. Additionally, audio-recordings were transcribed and analyzed with Thematic Analysis. In this procedure, four cards were categorized as high-rated: detailed explanations of performance class, detailed explanations of the influence of training focus (strength or cardio), detailed explanations of the influence of workout times, and detailed explanations of the influence of training cycles. Hence, the corresponding key components were added to the already correct key components in the user mental model from stage B to build the target mental model.

How to Explain: (D) Iterative Prototyping

Based on the *key components* from the *target mental model*, the *transparency team* and *employees* are co-designing prototype versions of the new UI. This stage is likely to take several iterations. (1) The *transparency team* and *employees* first identify possible locations for explanations in the current UI and the user flow. (2) The *transparency team* explores visualization possibilities for the explanations. These visualizations take the current UI design, the screen estate, corporate and system design guidelines as well as best practices in UI design into account.

Guiding Questions

How can the *target mental model* be reached through UI design? How and where can transparency be integrated into the UI of the system?

Why is this Important?

Since each system has a different UI design and a different user flow, we argue that is important to find visualizations that integrate explanations according to the look-and-feel of the UI in place and the available screen estate. Moreover, there are currently few guidelines in prior work and no best practices concerning the presentation format of explanations yet (see Background section), so that companies have to explore a way to visualize explanations.

Outcome
The outcome of this stage should be one or – ideally – several prototypes that explore possibilities of integrating explanations into the existing UI.

Exemplary Methods

- Focus groups, workshops and brainstorming sessions in the *transparency team* and with *employees* (approach taken)
- Design guidelines and best practices in UI design (approach taken)
- Low and high fidelity prototyping (approach taken)

Application to our Case
The first step in the iterative prototype design was a "How might we" brainstorming session [10] based on the target mental model in the transparency team and with other employees. In this process, all team members quietly sketched out ideas how and when key components of the target mental model could be visualized in the current UI and user flow and subsequently explained their ideas to the rest of the team. To select the most promising ideas, all team members then voted for five ideas perceived as most promising followed by a short discussion of pros and cons. The workshops allowed the transparency team to generate a list of promising implementation ideas, while leveraging the expertise of employees and adhering to the the design requirements and guidelines of Freeletics.

Next, we implemented two of the most promising ideas in a series of low and high fidelity prototypes. For this purpose, Freeletics had provided us with the necessary documents about the current user flow in the app. Within this user flow, we identified suitable locations for the integration of transparency into the Freeletics Bodyweight Coach UI.

The prototypes were evaluated and subsequently refined in several informal user testings (with only one or two participants), design workshops and expert reviews with the UI design experts within the transparency team and with other employees. As a result, we designed and developed two high-fidelity click prototypes that followed the corporate design guidelines of Freeletics and best practices in UI design. These click prototypes simulated the normal user flow as far as necessary in order to integrate the new transparency design concepts. Explanations were embedded in the user flow following an on-demand approach recommended by prior work [21, 22]: Users were first presented with an explanation of low detail and could get more detailed information in another screen if they wanted to know more.

How to Explain: (E) Design Evaluation

This stage completes the design process and aims at evaluating the effectiveness of the prototype(s) developed in stage D. Effectiveness is measured in terms of the differences between the *user mental model* and the *target mental model*. An effective prototype will reduce the differences between the *user mental model* and the *target mental model*.

(1) The *transparency team* elicits the *user mental model* based on the new UI. Again, the *transparency team* needs to decide whether to recruit the same or different users as in stage B (within-group or between-groups experimental design) according to resources and constraints. However, in this stage, the user group should not be the same as in stage C, where the *target mental model* was elicited, as these participants have been exposed to explanations before. For the same reason, if multiple prototypes have been developed (as in our project), it is sensible to recruit different user groups for each prototype to investigate the effect of the different design solutions on users' mental models (in contrast to the general recommendation to show users more than one prototype to elicit their opinion [44]). (2) The *key components* of the new *user mental model* are compared to those of the *target mental model*. The *transparency team* may decide to go back to stage D if the prototype is not effective enough, that is, if *user mental model* and *target mental model* do not match.

Guiding Questions
How has the *user mental model* developed? Has the *target mental model* been reached?

Why is this Important?
The final stage tests whether the prototype design has reached the intended goal of improving users' mental models.

Outcome
The outcome of this stage are the learnings from the evaluation of the prototype(s) that describe which design changes improved users' mental models and should therefore be integrated into the working system.

Exemplary Methods

- Interviews [7] (approach taken)
- Data collection in problem-solving tasks [32] (approach taken)
- Hypothetical scenarios [22]
- Think-Aloud [31] (approach taken)
- Questionnaire [21]

Application to our Case
Our prototypes were evaluated with seven and nine active Freeletics users, respectively, six women and nine men between 19 and 47 (*M*=30.7 years). Participants were randomly assigned to one of the two prototypes and guided through the prototype by the user flow that they knew from using the app. We encouraged them to explore the prototypes and think-aloud (1) what they saw on each screen, (2) what they thought could be done, and (3) where they would tap next.

Participants were then asked a set of comprehension questions about the key components explained in the prototype in a semi-structured interview, for example, "May some of your workouts be longer or more intense compared to other Freeletics users of the same sex and age? If so, why?", or "What influence do better or worse workout times have?". We also wanted to know whether participants perceived the explanations as useful and interesting, and whether they liked the level of detail or had rather wanted more or less information. We audio-recorded all answers and used screen-recording to capture their behavior while interacting with the prototypes. In the last part of the study, participants filled out the System Usability Scale [4] for the prototype they had tested.

Answers to the semi-structured interview questions were transcribed and then compared to model answers we had prepared beforehand. We rated given answers as “correct”, “rather correct”, “rather wrong”, and “wrong”. We found that overall, participants’ answers indicated “correct” or “rather correct” understanding of the training cycles (15 participants), workout times (all participants) and performance classes (14 participants). For example, they could explain the purpose of training cycles and knew where they could get information about how intense and long workouts in the next week will be in the prototype. Moreover, 14 participants perceived the explanations as interesting and informative and only two participants stated that they were rather not interested in the reasoning of the system. The prototypes reached an overall score of 89.6 and 82.0, respectively, on the System Usability Scale, which corresponds to “excellent” usability [4].

While this indicates the effectiveness of our prototypes in terms of the *target mental model*, we found room for improvement regarding the location of the explanations. For example, screen-recordings showed that buttons and links that would have led to more detailed on-demand explanations were completely overlooked by several participants.

In summary, the evaluation in this stage allowed us to demonstrate the effectiveness of the developed prototypes to all stakeholders while also highlighting directions for future work and opportunities for improvement.

LIMITATIONS

The presented process evolved over six months and has proven successful in the Freeletics design scenario. Yet, we acknowledge that it is certainly not the only, but a promising way to address the complexity of designing for transparency in real-world design settings. We will further elaborate on this point in the discussion section of this paper. Moreover, the presented methods are based on those used in the reviewed literature and should facilitate application of our process. They served well in answering the guiding questions in each process stage in our project – however, other participatory methods may be equally suitable.

Moreover, our process inherits the limitations that come with investigating mental models in general. Researchers’ opinions are divided when it comes to the elicitation of mental models: While Norman [32], for example, claims that asking users about their mental model is less reliable than collecting data in problem-solving tasks, Nielsen [31] states that methods like card sorting or think-aloud are valid approaches to elicit users’ mental models.

DISCUSSION AND FUTURE WORK

We have presented a stage-based design process to facilitate the integration of transparency in complex real-world design scenarios, which was developed and tested in an exemplary case with the Freeletics Bodyweight Coach. In the following sections, we reflect on our process and different views on the design of transparent systems.

Validity and Universality of our Process

Our process successfully addressed the challenges of the Freeletics project, met the different needs of the stakeholders involved, and improved users’ mental model of the Freeletics Bodyweight Coach. These points strongly indicate the validity of the process in the given scenario. Moreover, mental models as a system-indepentent concept have been used in prior work as a way to approach the design of transparent intelligent systems in different contexts [22, 23, 24]. We therefore assume that our process is valid beyond the scope of our case and universal in the sense that it is applicable to other, similarly complex scenarios involving other types of systems, yet concrete enough to serve practitioners as a starting point for their work. However, this needs further evaluation in comparable real-world settings, and we acknowledge that variants of the process may exist that are better suited for specific scenarios. We therefore invite other researchers to validate and build on our process in the future.

Possible Improvements

We acknowledge that even though our process aims at facilitating the integration of transparency, it still requires more effort in terms of cost and time than automatically generating explanations, which not all companies might be willing to invest. However, evaluation of the prototypes showed that users were able to improve their mental models based on the explanations resulting from our process. This strongly suggests that explanations tailored to the users’ interests may add to the value of an application by supporting appropriate usage while taking the companies’ needs into account. A point worth investigating in this context would therefore be whether the stages of our process could be fully or partially automated in the future to reduce the effort necessary to generate suitable explanations.

While prior work argues that both communications from the systems to the user and from the user to the system are important, for example for correcting system mistakes [7], the focus in our case, the Freeletics project, was on transparency of the system only. Although considering user-to-system communication was out of scope of the presented project, it is an important aspect for shaping interaction with intelligent systems in the future. We assume that our process might be equally helpful for designing for this direction of the communication between system and user, since the setting does not change and the resulting difficulties as well as the importance of mental models therefore might be comparable. However, this assumption needs to be validated in future work.

Another concept closely linked to transparency is trust in an intelligent system [41], which is of crucial importance for safety-related systems, for example, airplanes [29], but does also influence user satisfaction with systems such as recommenders [43]. Trust was not assessed during our project, but future work might investigate how our process and improved mental models affect trust, using frameworks such as [17].

Transparency Normativism vs Pragmatism

Transparency in HCI is a fuzzy and multi-facetted concept which spans a variety of research areas. Explaining how a system works might enable users to correct system decisions through feedback (as in [21, 23]), might improve understanding of the system recommendations and thus foster efficient or effective use [35, 36], prevent mis- and disuse of a system through increased trust [39] or protect users' privacy [19] , to mention a few exemplary purposes. When following the call for transparency in intelligent systems, one therefore has to wonder what the standard for transparency should look like – what is a "genuinely" transparent system? Some underlying algorithms such as decision trees and Naive Bayes might fairly well be made transparent, while others, such as neural networks, are black-boxes by nature, and inherently not interpretable even by their developers. Hildebrandt [16] argues that "decisions that seriously affect individuals' capabilities must be constructed in ways that are comprehensible as well as contestable. If that is not possible, [...] such decisions are unlawful". As a consequence, she claims that intelligent algorithms that cannot explain themselves to users must not be deployed. However, this view would imply that undoubtedly useful machine learning algorithms, in particular deep learning algorithms, such as used in computer vision and self-driving cars, were not to be applied in practice as researchers are still working on ways to make their decisions explainable [2]. A normative view on transparency might also imply that any kind of abstraction from the algorithms' reasoning is not consistent with the principle of transparency. This might nudge companies to present users with extensive legal declarations, similar to End User License Agreements [14], and thus undermine the "right to explanation" emerging through EU's GDPR [33] with information that requires a lot of effort to process.

In this discussion, we follow a pragmatic view on transparency that sees transparency as a trade-off between practical applicability and literal transparency – transparency "as good as possible". We argue that transparency should be designed in a way that benefits users most while taking requirements of other stakeholders into account.

A pragmatic view also acknowledges that users might not be interested in *all* of the systems underlying reasoning [25], and that they find some pieces of information more interesting than others [25] (as it was the case in our project with Freeletics). Based on our learnings in this case study, we hypothesize that users might follow the principle of *satisficing* [42] also in the context of transparency: Transparency satisfices users if it allows them to build mental models good enough to predict and explain the observed system behavior, even though the given explanations might not include all factors relevant for the workings of the algorithm. In this paper, we presented a process that aims at balancing conflicting influencing factors when designing for transparency and, hence, to help designing for satisficing. Once this approach has been applied to a variety of scenarios, future work might be able to extract general patterns or guidelines for transparency satisficing.

Establishing Best Practices for Transparency Design

Our work raises the questions if it is possible to establish best practices for transparency design similar to best practices in the design of UIs or user interaction, and if so, how they can emerge. These questions will become more urgent as companies need to fulfill the legal requirements of the European Union's General Data Protection Regulation. Gregor and Benbasat [13], for example, claim that the design of explanations in intelligent systems should be theory-based instead of bottom-up. From our experience with the Freeletics case however, where we faced considerable difficulty of working top-down from existing guidelines, we argue that this top-down approach needs to be complemented with a bottom-up approach that focuses on extracting learnings on a case-by-case basis. These learnings from practice could then be condensed into higher-level guidelines, or design patterns, which might possibly focus on specific product domains that share similar characteristics and challenges, or similar goals.

CONCLUSION

In this paper, we suggested a stage-based, participatory design process to help designers to integrate transparency into applications in complex real-world scenarios involving multiple stakeholders. This process was developed and validated during a six months project in cooperation with Freeletics, in which we worked on integrating transparency into the Freeletics Bodyweight Coach, an intelligent fitness application whose AI calculates personalized exercising plans for the users. The process makes use of transparency as a means to improve users' mental models. Our approach follows a pragmatic view on transparency, where transparency is implemented based on (1) what can generally be made transparent of the underlying algorithm, and (2) the information that is interesting and relevant to users. We argue that designing UIs for transparency may greatly benefit from case-by-case learning, and hope that our process can pave the way for more transparency in similarly complex design settings.

ACKNOWLEDGMENTS

We are greatly indebted to Natalia Patiño Pérez, Renato Pereira and Peter Just, who were an amazing help during discussion, prototype design and user research. We also thank Laith Alkurdi for the workshop on the AI of the Freeletics Bodyweight Coach, and Robert Schmidt and Tobias Seitz as well as the anonymous reviewers for their helpful comments on earlier versions of this manuscript.

REFERENCES

1. Victoria Bellotti and Keith Edwards. 2001. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212.

2. Mariusz Bojarski, Philip Yeres, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Lawrence D. Jackel, and Urs Muller. 2017. Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car. *CoRR* abs/1704.07911 (2017). `http://arxiv.org/abs/1704.07911`

3. Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. DOI: http://dx.doi.org/10.1191/1478088706qp063oa

4. John Brooke and others. 1996. SUS – A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

5. ACM U.S. Public Policy Council and ACM Europe Policy Committee. 2017. Statement on Algorithmic Transparency and Accountability. (2017). https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

6. Kenneth James Williams Craik. 1967. *The Nature of Explanation*. Vol. 445. CUP Archive.

7. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The Effects of Transparency on Trust in and Acceptance of a Content-Based Art Recommender. *User Modeling and User-Adapted Interaction* 18, 5 (20 Aug 2008), 455. DOI: http://dx.doi.org/10.1007/s11257-008-9051-3

8. Jérôme Dinet and Muneo Kitajima. 2011. "Draw Me the Web": Impact of Mental Model of the Web on Information Search Performance of Young Users. In *Proceedings of the 23rd Conference on L'Interaction Homme-Machine (IHM '11)*. ACM, New York, NY, USA, Article 3, 7 pages. DOI: http://dx.doi.org/10.1145/2044354.2044358

9. Paul Dourish, Annette Adler, and Brian Cantwell Smith. 1996. Organising User Interfaces around Reflective Accounts. *Reflection, San Francisco, CA* (1996).

10. Stanford d.school. 2017. "How Might We" Questions. (2017). Retrieved 07 October 2017 from https://dschool.stanford.edu/resources/how-might-we-questions

11. Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718.

12. Freeletics. 2017. Freeletics Bodyweight Coach. (2017). Retrieved 20 December 2017 from freeletics.com/en/bodyweight/coach/get

13. Shirley Gregor and Izak Benbasat. 1999. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly* (1999), 497–530.

14. Jens Grossklags and Nathan Good. 2007. Empirical Studies on Software Notices to Inform Policy Makers and Usability Designers. *Financial Cryptography and Data Security* (2007), 341–355.

15. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, New York, NY, USA, 241–250. DOI: http://dx.doi.org/10.1145/358916.358995

16. Mireille Hildebrandt. 2016. *The New Imbroglio. Living with Machine Algorithms*. Amsterdam University Press, 55–60.

17. Kimberly F. Jackson, Zahar Prasov, Emily C. Vincent, and Eric M. Jones. 2016. A Heuristic Based Framework for Improving Design of Unmanned Systems by Quantifying and Assessing Operator Trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, 1 (2016), 1696–1700. DOI: http://dx.doi.org/10.1177/1541931213601390

18. Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. 2015. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, Ottawa, 39–52. https://www.usenix.org/conference/soups2015/proceedings/presentation/kang

19. Judy Kay and Bob Kummerfeld. 2013. Creating Personalized Systems That People Can Scrutinize and Control: Drivers, Principles and Experience. *ACM Trans. Interact. Intell. Syst.* 2, 4, Article 24 (Jan. 2013), 42 pages. DOI:http://dx.doi.org/10.1145/2395123.2395129

20. Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User Preferences for Hybrid Explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. ACM, New York, NY, USA, 84–88. DOI: http://dx.doi.org/10.1145/3109859.3109915

21. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 126–137. DOI: http://dx.doi.org/10.1145/2678025.2701399

22. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/2207676.2207678

23. Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In *Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing (VLHCC '10)*. IEEE Computer Society, Washington, DC, USA, 41–48. DOI: http://dx.doi.org/10.1109/VLHCC.2010.15

24. Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. DOI: http://dx.doi.org/10.1109/VLHCC.2013.6645235

25. Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing (UbiComp '09)*. ACM, New York, NY, USA, 195–204. DOI: http://dx.doi.org/10.1145/1620545.1620576

26. Brian Y. Lim and Anind K. Dey. 2010. Toolkit to Support Intelligibility in Context-aware Applications. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*. ACM, New York, NY, USA, 13–22. DOI: http://dx.doi.org/10.1145/1864349.1864353

27. Brian Y. Lim and Anind K. Dey. 2011. Design of an Intelligible Mobile Context-aware Application. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. ACM, New York, NY, USA, 157–166. DOI: http://dx.doi.org/10.1145/2037373.2037399

28. Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2119–2128. DOI: http://dx.doi.org/10.1145/1518701.1519023

29. Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. Shaping Trust through Transparent Design: Theoretical and Experimental Guidelines. In *Advances in Human Factors in Robots and Unmanned Systems*. Springer, 127–136.

30. Neville Moray. 1987. Intelligent Aids, Mental Models, and the Theory of Machines. *International Journal of Man-Machine Studies* 27, 5 (1987), 619 – 629. DOI: http://dx.doi.org/https://doi.org/10.1016/S0020-7373(87)80020-2

31. Jakob Nielsen. 2010. Mental Models. (2010). Retrieved 04 October 2017 from www.nngroup.com/articles/mental-models/

32. Donald A. Norman. 1983. Some Observations on Mental Models. *Mental Models* 7, 112 (1983), 7–14.

33. PrivacyPlan. 2017a. Article 13 EU GDPR "Information to be Provided where Personal Data are Collected from the Data Subject". (2017). Retrieved 27 September 2017 from http://www.privacy-regulation.eu/en/13.htm

34. PrivacyPlan. 2017b. Article 22 EU GDPR "Automated Individual Decision Making, Including Profiling". (2017). Retrieved 27 September 2017 from http://www.privacy-regulation.eu/en/22.htm

35. Pearl Pu and Li Chen. 2006. Trust Building with Explanation Interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI '06)*. ACM, New York, NY, USA, 93–100. DOI: http://dx.doi.org/10.1145/1111449.1111475

36. Pearl Pu and Li Chen. 2007. Trust-Inspiring Explanation Interfaces for Recommender Systems. *Knowledge-Based Systems* 20, 6 (2007), 542 – 556. DOI: http://dx.doi.org/https://doi.org/10.1016/j.knosys.2007.04.004 Special Issue On Intelligent User Interfaces.

37. Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating Recommender Systems from the User's Perspective: Survey of the State of the Art. *User Modeling and User-Adapted Interaction* 22, 4 (01 Oct 2012), 317–355. DOI:http://dx.doi.org/10.1007/s11257-011-9115-7

38. Virpi Roto, Heli Väätäjä, Satu Jumisko-Pyykkö, and Kaisa Väänänen-Vainio-Mattila. 2011. Best Practices for Capturing Context in User Experience Studies in the Wild. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. ACM, 91–98.

39. Garrett Sadler, Henri Battiste, Nhut Ho, Lauren Hoffmann, Walter Johnson, Robert Shively, Joseph Lyons, and David Smith. 2016. Effects of Transparency on Pilot Trust and Agreement in the Autonomous Constrained Flight Planner. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. 1–9. DOI: http://dx.doi.org/10.1109/DASC.2016.7777998

40. James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek Abdelzaher, and John O'Donovan. 2015. Getting the Message? A Study of Explanation Interfaces for Microblog Data Analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 345–356. DOI: http://dx.doi.org/10.1145/2678025.2701406

41. Thomas B. Sheridan. 1992. *Telerobotics, Automation, and Human Supervisory Control*. MIT press.

42. Herbert A. Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (1955), 99–118. DOI:http://dx.doi.org/10.2307/1884852

43. Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. ACM, New York, NY, USA, 830–831. DOI:http://dx.doi.org/10.1145/506443.506619

44. Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the Right Design and the Design Right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 1243–1252. DOI: http://dx.doi.org/10.1145/1124772.1124960

45. Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How It Works: A Field Study of Non-technical Users Interacting with an Intelligent System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 31–40. DOI: http://dx.doi.org/10.1145/1240624.1240630

46. Arnold P. O. S. Vermeeren, Effie L-C. Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. User Experience Evaluation Methods: Current State and Development Needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. ACM, 521–530.

47. Jed R. Wood and Larry E. Wood. 2008. Card Sorting: Current Practices and Beyond. *Journal of Usability Studies* 4, 1 (Nov. 2008), 1–6. http://dl.acm.org/citation.cfm?id=2835577.2835578