aerospace

MDPI

Article

Playful Probing: Towards Understanding the Interaction with Machine Learning in the Design of Maintenance Planning Tools

Jorge Ribeiro * and Licínio Roque

CISUC—Centre Informatics and Systems, Informatics Engineering Department, University of Coimbra, 3004-531 Coimbra, Portugal
* Correspondence: jamr@dei.uc.pt

Abstract: In the context of understanding interaction with artificial intelligence algorithms in a decision support system, this study addresses the use of a playful probe as a potential speculative design approach. We describe the process of researching a new machine learning (ML)-based planning tool for maintenance based on aircraft conditions and the challenge of investigating how playful probes can enable end-user participation during the process of design. Using a design science research approach, we designed a playful probe protocol and materials and evaluated results by running a participatory design workshop. With this approach, participants facilitated speculative design insights into understandable interactions, especially with ML interaction. The article contributes with a design of a playful probe exercise to collaboratively study the adjustment of practices for CBM and a set of concrete insights on understandable interactions with CBM.

Keywords: playful probing; cultural probes; design requirements; games research; participatory design; decision support systems; condition-based maintenance; machine learning interaction

Citation: Ribeiro, J.; Roque, L. Playful Probing: Towards Understanding the Interaction with Machine Learning in the Design of Maintenance Planning Tools. *Aerospace* **2022**, *9*, 754. https://doi.org/10.3390/aerospace9120754

Academic Editors: Bruno F. Santos, Theodoros H. Loutas and Dimitrios Zarouchas

Received: 31 October 2022
Accepted: 23 November 2022
Published: 26 November 2022

1. Introduction

There is increasing demand to incorporate powerful AI (artificial intelligence) algorithms in information systems, often in data sensitive or critical decision support systems. Introducing these algorithms in a critical and highly regulated operational context resists experimentation and raises new design challenges.

Such a scenario of speculative design raises challenges regarding design approaches that make it easier to understand new interactions, together with the design and human appropriation and control over new AI tools. In this context, we cannot perform ethnography of as yet nonexistent interactions; neither can we apply conventional requirements eliciting approaches that presume such knowledge.

Moreover, interdependence between new tool development and new practices in a strongly regulated context inhibits experimentation, creating a cultural deadlock. Such a scenario demands a generative approach informed by current practices and also new AI possibilities as they become available through research. Instead, a participatory approach is needed to empower practitioners [1] to develop new ways of working with and designing AI-enhanced decision support systems. However, it remains unanswered which participatory approach is best suited to the design of the CBM planning tool.

Cultural probes, as proposed by Gaver et al., are "an approach of user-centred design for understanding human phenomena and exploring design opportunities" [2] focused on new understandings of technology. Cultural probes do not give comprehensive information about people and their practices; rather, they provide fragmentary clues about their lives and thoughts [3]. The researcher or the designer has the mission of putting together the pieces of that puzzle and making findings emerge from that. Therefore, cultural probes can be a tool for designers to understand users [4]. F. Lange-Nielsen show some studies in which probes are used as a scientific method or a design tool [5], and Hutchinson et al.

show how technology probes can be a promising new design tool in the design of new technologies [6]. Vasconcelos and others report on a study inspired by the concept of culture probes describing the process of creating a low-, medium-, and high-fidelity prototypes for a cognitive computer game [7].

J. Wallace et al. argue that design probes provide more than just inspiration for design and can be used to mediate both the relationship between participant and researcher and her own feelings [8]. However, we need more than mediation: we need a way to gather insights on how to use AI to plan maintenance. The role of play in cultural development has been recognized at least since Huizinga [9]. There has been extensive work on this topic in the scientific community, as discussed in Section 2.2.

Adopting the concept of playful probes could potentially enable the exploration of AI/ML methods by helping to develop the participants' perspective and the appropriation of new tools [10]. However, we still lack understanding on a research question: *How can we explore playful probing to draw insights into understandable interactions with AI/ML tools?*

The purpose of the design case we present is to obtain insights and better understand how condition-based maintenance (CBM) planning can be introduced in a critical operations sector: aircraft maintenance. More specifically, we aim to gain insights on understandable interactions of how to perform aircraft maintenance planning assisted by a machine learning agent. In this paper, we report on a design science research process that runs a participatory playful probing workshop for evaluating a proposed design supported in a virtual paper prototype. This was done while simulating an ML-enhanced CBM maintenance context with aircraft maintenance domain experts. Videography, dialogues, and interviews were *open coded* for content analysis and summary of design insights towards the proposed research question.

This study contributes to the design of a playful probe exercise to collaboratively study the adjustment of practices for CBM. As a result of this exercise, a set of eleven concrete insights about understandable interactions with CBM maintenance planning emerged. In this essay, we present and discuss these understandable interactions.

The next section refers to some background concepts related to aircraft maintenance and cultural probes. Then, we describe the initial exploration of the work, followed by the process of playful probes that explains how the various methods and procedures are integrated. Subsequently, we provide the collection and analysis of data obtained in the workshop, followed by the discussion and synthesis of the understandable interactions. The final part includes the conclusions of this study.

2. Background and Related Work

In this chapter, we present the literature on the current state of aircraft condition-based maintenance, followed by cultural probing and the evolution to playful probing.

2.1. Aircraft Condition-Based Maintenance

Current aircraft maintenance (AM) is based on the task-oriented MSG-3 model defined in 1979 by the Air Transport Association (ATA) [11]. The MSG-3 method defines the obligation to carry out scheduled and routine maintenance in a given structure, but it also allows slack for unscheduled or non-routine maintenance that results in maintenance actions to correct divergences detected during scheduled maintenance tasks. The AM domain poses new challenges for the design of decision support systems where a human and machine learning (ML) confluence can open new opportunities such as a better approach to perform maintenance in the aircraft maintenance industry: (CBM) [12]. This technique exploits ML-based components and systems failure forecasts to schedule maintenance at the most opportune moment instead of using a fixed interval approach, increasing aircraft availability and safety while reducing costs [13]. CBM is being increasingly adopted, including these ML processes that produce remaining useful life (RUL) estimates for aircraft system components and generate updated plan proposals for user validation.

However, such a critical operational context is highly regulated and resists experimentation. This scenario raises new and relevant challenges for design approaches that can enable evolution of current practices in the field by designing for human appropriation and control over new ML algorithms. Under such a context, we cannot perform ethnography of as yet nonexistent practices; neither can we apply a priori approaches for eliciting requirements. Instead, a participatory approach [1] is needed to empower practitioners to understand and develop new ways to work with and design ML-enhanced decision support tools.

2.2. Playful Probing

The role of play in cultural development has been recognized at least since Huizinga [9]. A playful probing approach uses games designed specifically for the study and tailored to the research area and purpose of the study [14]. Research [15] suggests that a game designed for playful probing "opens up for a playful and autonomous environment for data-gathering which involve learning about individual and shared social practices". The playful probes technique uses similar principles to those of cultural probes while exploiting games as a research tool to enable learning and data collection. Through the use of support artefacts, cultural probes allow participants to document their activities and experiences to be used as research material. This allows collection of the participants' perspectives in the process, allowing them to explore new things beyond the expected.

As defined by Huizinga [9], play is an experience outside the real world, *a magic circle* where the player can explore, experiment, and provoke in a safe environment. With Huizinga, we learned to recognize the role of play in cultural development. Playful probes might enable a planner to enter this bubble and establish the dialectic with a new AI approach and well-defined processes. As a result, this might allow us to identify, anticipate, and understand possible problems. The concept of play has been addressed before [3–6,8] in the context of studying novel interaction design proposals. Playful probing [14] espouses similar principles of cultural probes while exploiting games as a research tool to enable learning and data collection [7,15–17]. However, we still must identify how to research risk scenarios. The answer can be a simulation game as an enabler of participatory context. However, we do not know how we can create such a participatory context using a simulation game.

Since the development of modern digital computers, computer simulations have been used for modelling and studying systems [18]. Simulation games have been used to formalize scientific problems and have also been adopted by academia in game form as ways to formalize and study, e.g., economic and social behavioural phenomena. Simulation games can make it possible to create and study scenarios without compromising actual maintenance operations. Turning simulations into games can also enable the exploration of behaviour in alternative settings.

Previous research [19] showed that playful probing artefacts can be used to design new ML algorithms in a critical and highly regulated operational context, and another preview study [20] showed how to elicit ideas for integration of maintenance planning practices with ML estimation tools and the ML agent using playful probes. However, we still do not know how we can use a simulation game in a participatory context to allow us to study insights about understandable interactions with AI/ML decision support tools.

3. Initial Exploration of Work

In this chapter, we describe our initial exploration of the topic. Prior to the participatory process, we performed bibliographic research of the state-of-the-art in human–ML interaction and aircraft maintenance. After this initial exploration, the literature review was deepened and directed to the research question of this study, which has already been presented in Section 2.

3.1. Human–ML Interaction for CBM

A better human–computer confluence can be achieved by enabling co-creation between the user and the artificial intelligence (AI) algorithm and putting explainability at the core of user autonomy and empowerment. Some studies [21,22] provide human–AI interaction guidelines, but explainable, accountable, and intelligible systems remain key challenges [23–25]. While progress has been made in explainability and interpretability [26–28], the design of AI interfaces focusing on co-creativity remains a challenge.

Bødker et al. [29] points to the problem of appropriation and control as people learn new technologies and update cooperative work practices. We need to identify how a tool can be created to support a dialogue between the planner and an ML algorithm while also preserving the autonomy and control of the human agent in a risky context. A process is needed to design a new tool enabling a new meaningful simulated practice.

Simulations and games have been used to model and study systems [18], formalise scientific problems, and study economic and behavioural phenomena [30]. Simulations allow us to test and explore new interaction approaches, while new practices can emerge in a meaningful but simulated context. Playful probes can be the variation of the cultural probes approach [10] that allows learning and data collection [9] in a simulated playful environment.

3.2. Maintenance Planning

It is not possible to find technical details in the literature of how a company such as the one we are studying performs maintenance planning. To obtain valuable information about maintenance practices, teams, and tools, we applied the PD methods described in the next section.

Semi-structured interviews were performed by the two authors of this paper with two maintenance planning workers: one with a lot of experience and responsibilities in the domain, and the other with a few years of experience. The interview was recorded, transcribed, and analysed a posteriori. It served, to a large extent, to enlighten us on how maintenance is done in everyday life, the volume and type of daily work, how the maintainers cooperate with the other teams, and how the maintainers achieve their work using their specific tools. Throughout the interview, speculative questions were asked to gain understanding on whether indicators could be used for a CBM paradigm shift. At this stage, we did not give relevance to how ML algorithms (in this case an ML agent) could help in the daily work of planners.

A guided visit to the facilities took place at the maintenance planning building, maintenance control centre, and hangars. It was important to perceive the skills, tools, and particularities of each team in place, particularly among maintenance planners and maintenance engineers.

The information was also complemented with some presentations on the project setting and the sharing several technical documents with maintenance-related details.

Based on this information, we were able to synthesise the main concepts that allow us to represent how maintenance is performed:

- Block: predefined routine maintenance, usually heavy and with due dates (as "A-checks").
- Cluster: usually a flexible, small group of tasks that can be routine or non-routine, such as reactive or preventive maintenance; can have due dates, RUL, both, or none.
- Flight: aircraft movement between airports. It is not possible to do any maintenance to the aircraft in this period.
- Hangar: place where maintenance is performed. It has several restrictions, such as time, materials, and labour.

The flight element is not currently viewed by maintenance planners in planning software. However, we consider it relevant to include it in this study, as it has a relevant impact on hangar maintenance and resources. This information is important to discovering

the process of playful probes, described in the next section, especially for the creation of the playful probe paper prototype used in the workshop.

4. The Process of Playful Probes

In this section, we describe how the playful probing process was set up. Bearing in mind that we intend to speculatively study a CBM maintenance planning tool, we proposed a combination of methods for the discovery process: a cooperative future workshop, focus groups, and playful probes using a paper prototype. The exercise included: in a first phase, a cooperative future workshop using playful probes, followed by a focus group functioning as a reflection and exploitation of what happened during the first phase. The focus group was a guided discussion followed by a structured interview sent by email after workshop analysis.

We designed playful scenarios and materials as well as a playful probing protocol. In the process, we expected to open new relevant questions about ML-based RUL estimators, maintenance planning practices, and how to design human–tool interaction in a future computer prototype.

For relevance and simplicity, we determined that in this study we would just play with shorter maintenance work cycles, called "A-checks", using one RUL indicator for each aircraft maintenance package instead of representing an RUL for each component and system inside that maintenance.

This study was conducted as part of an ongoing design science research (DSR) approach [31], and it was conducted in a workshop session. Given the schedule restrictions and the difficulty in obtaining the agendas of our project partners, we accepted that the institution/company would recruit the participants.

The workshop was carried out with two researchers and two domain experts in maintenance management (hereinafter designated P1 and P2), both male and between 20 and 40 years old, with strong backgrounds in aviation and practical knowledge of planning tools. They played the simulation scenario and cooperated to solve each maintenance problem presented, reflecting on the mediating role of the new ML tools. The workshop was facilitated by the researcher—who ensured the application of the protocol and clarification of doubts on play scenarios and materials, e.g., role playing the gamemaster role—and the designer researcher, who assisted in the discussion.

The next subsection describes the preparation of playful probe material and how it was applied.

4.1. Playful Probe Preparation

Given the focus of developing a new CBM planning tool with the concept of short-term maintenance, we chose to simplify the concepts that allow us to represent and explore how short-term maintenance is performed.

We created a simplified and simulated version of a current standard planning activity to speculate on how evolution can be done with CBM. However, we considered an exception: Instead of preprocessing the fleet information to find maintenance opportunities, as is the current practice, we decided to include the flight plan and consider any time when the plane is not being used as a potential period to perform maintenance. We also decided not to include any interaction with the ML planning agent so as not to bias the discussion and solutions presented by the participants. The materials were created with shapes and colours to be easily identifiable and distinguishable. We wanted objects to be easily playable and studied and focused on functionality, manipulability, and the resulting discourse rather than aesthetics.

The steps carried out in this study are presented below.

1. **Materials design**
   The materials presented in Figure 1 were based on the main maintenance concepts set out in Section 3:
   - Row: one row represents an aircraft.

- Column: representation of time. One column represents one day.
- Flight: blue ribbons represent aircraft flights in the respective aircraft row.
- Registration time: the limit time to register the aircraft to some maintenance slot is 30 days.
- Open time: the limit time to open a new workscope (create new maintenance) is 21 days.
- Block: red rectangles represent predefined routine maintenance (with due date from the maintenance planning document). When moved before the registration limit, it must be registered as a new block after this registration limit.
- Cluster: group of tasks that represent other types of A-checks (small maintenance) with due date, RUL, both, or none. If a cluster is moved, it must be moved after the open workscope limit unless it is joined to a block.

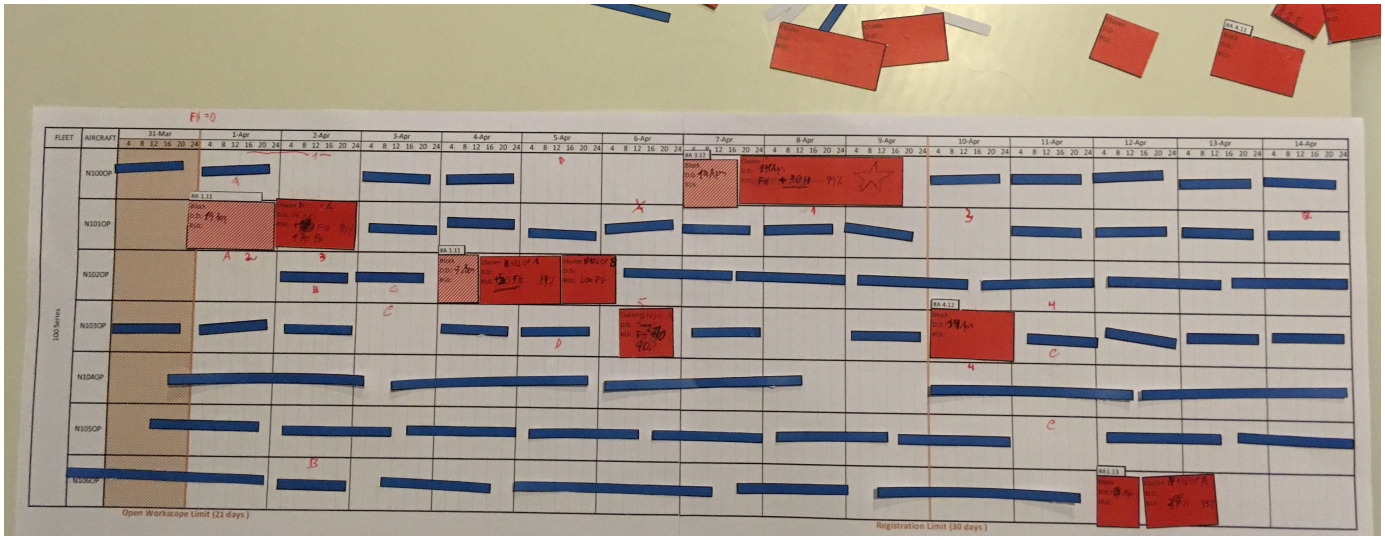

Figure 1. Paper prototype materials printed for testing.

In this phase of the discovery process, we chose to create materials that tried to faithfully represent current maintenance concepts as a starting point. We decided to use only a small speculative detail of CBM maintenance in these materials: the RUL indicator (in flight hours) that was included in some clusters.

2. **Resolution path**
To enable this task in a limited time, we created a structured resolution path. The beginning of the resolution was linear and could only progress one way. Participants faced the simplest concepts of flight planing and maintenance. Subsequently, the resolution would lead to a path where users would necessarily be faced with more complex issues such as conflicting conditions and 90% confidence RULs.
This probe was designed so that during the rehearsal, the disposition of visual artefacts confront participants with situations that can lead to debate and the generation of insights. The main ones were:

   - Introducing block and cluster grouping—Is it possible to group all the maintenance in these two typologies? How do we deal with the deadlines of each type?
   - Introducing estimates with 90% confidence—Does it make sense to have a large degree of uncertainty? How do we represent it to enable decisions?

3. **Material digitalization**
To prepare the virtual workshop, all artefacts were designed digitally but printed and rehearsed with manually as is common in paper prototype exercises (Figure 1). After testing multiple approaches to instrument the playful probing with visual

artefacts, we adjusted size and complexity, and the exercise was migrated to the digital collaboration tool (Figure 2).

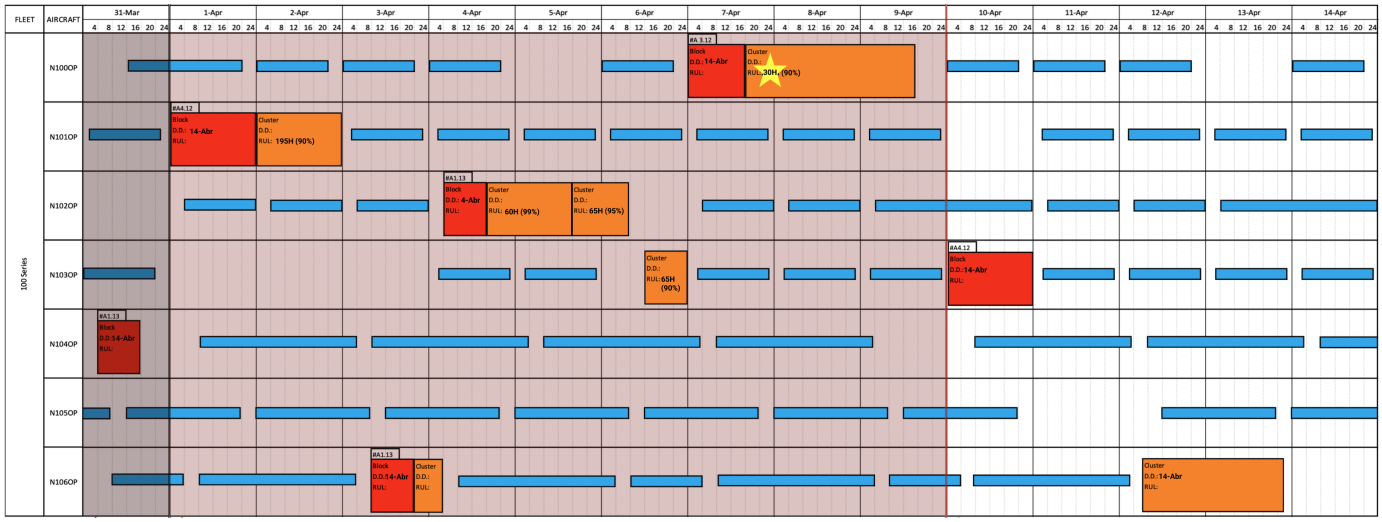

Figure 2. Maintenance scheduling problem presented to the participants in the experimental session.

Next, we describe the steps used in the rehearsal of the playful probing workshop.

4.2. Playful Probe Workshop

The workshop included, in a first phase of focus, a cooperative future workshop using playful probes (Steps 1 and 2), followed by a focus group functioning to reflect on and exploit what happened during the playful probe exercise through a guided discussion and an interview (Step 3).

4. **Briefing**
   In an initial part of the playful probing workshop, an introduction was made explaining the basic maintenance elements of the game and demonstrating how to solve a simple problem (Figure 3).

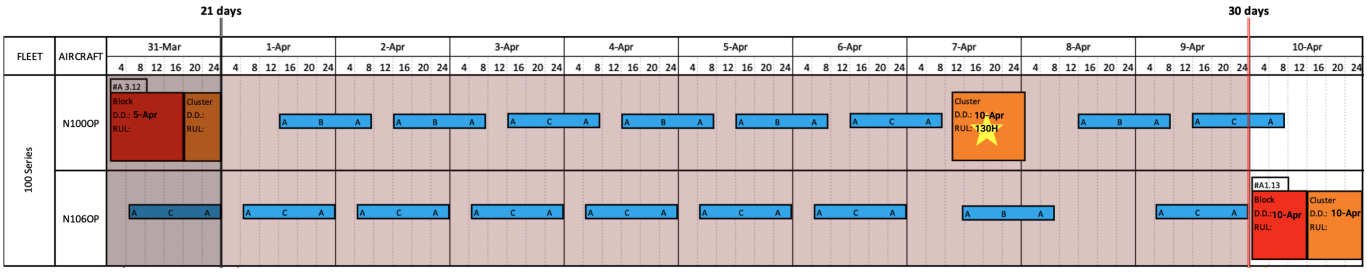

Figure 3. Minimal scheduling problem to introduce the rules and basic movements of the game.

The canvas represents a fleet of only two aircraft, with flights and maintenance distributed over time and using a minimum block time of 4 h. To simplify the maintenance problem for a first iteration of the game design, only three types of artefacts were created with which the participants could interact (drag and drop), representing the maintenance work of an "A-check". For simplicity, we assumed that there was only one hangar with a maintenance team available, so it was not possible to do multiple maintenance procedures at the same time. This part lasted around 10 min, and the participants cleared some doubts about the game but did not interact with artefacts.

5. **Running the playful probing participatory design workshop**
   In this part of the experimental session, artefacts were presented to participants with a non-trivial maintenance scheduling problem to be solved (Figure 2). The participants'

voices and the collaborative canvas were recorded while they presented their ideas and played with the representations to solve each maintenance problem. The facilitator acted as gamemaster, answered participants' questions about whether they could take some actions, alerted them when they were ignoring some important condition, and tried to get them to explore the problem boundaries in a dialogue with the material representations.

Exploration developed freely to solve each game problem, with no constraints regarding order, time, or the management of concurrency among open explorations; the facilitator favoured out-loud dialogue and the explicit manipulation of the representations as a form of dialogical imagination among participants. Given the habitual nature of play, we expected the emergence of self-directed and highly autonomous activities driven by participants' playful trajectories actively exploring the boundaries of the gameplay scenario.

6. **Debriefing debate**
Shortly after the participants solved the planning problem, the focus group occurred, in which a broader discussion space was opened to reflect on the current state of maintenance and how CBM can be used in the future.

4.3. After the Playful Probe Workshop

7. **Semi-structured email interview**
After reviewing the recordings, specific interview questions were sent to the participants with the intention of clarifying or deepening the reflections they expressed during the play and debriefing phases. The first group of questions focused on the experience and interpretation of the participants about the exercise. The leading questions were:

> How did you perceive the experience from the moment where the problem appeared with a yellow star to the reached solution? What did you find most challenging and why?

The second group were speculative questions about using an ML agent to help with maintenance planning. The leading questions were:

> How would you briefly narrate a planner using the ML planing/scheduling with this interface? At which moments ML should be called in to provide a new solution or a partial solution to the planner?

The third group of questions focused on the visualization, interpretation, and control of RUL indicators. The leading questions were:

> Did you experience difficulties visualizing/interpreting RUL indicators? Can you anticipate some improvement in the way we present information to give better control to the planner?

The last group of questions focused on the playful probing exercise itself. The leading questions were:

> What did you thought about the session technique used: should we make some changes? were the materials limiting in anyway that needs to be fixed? did it help generate or make explicit some insights about the subject mater?

The email interviews were typically an extrinsic reflection on the experience, a post-reflection. It will be discussed later in the discussion section.

8. **Data collection and in-depth content analysis**
The playful probing workshop generated audio and video recordings, dialogue text, and interview transcriptions. A video was made (with informed consent) of the conversation between the participants and the manipulation of game artefacts during the playful probe workshop. All video data were analysed (verbal and actions with materials) to generate initial codes. Then, the recording was split into 30 s segments and coded into groups. These grouping were based on the intrinsic self-analysis of

the experience that emerged in the conversation generated as participants played the scenario. Data collection and analysis of the workshop are described in the next section.

5. Data Collection and Analysis

The workshop recording was transcribed, and the content was analysed to draw main emergent categories. These classes were organized in a taxonomy that then served to guide the coding process. Subsequently, the recordings were divided into 30 s segments, and each segment was classified into one or more class. The two major areas were focus and reflection (Level 1). Focus represents when participants were focused on something directly related to the artefacts of the game; reflection marks when participants expressed some reflected thoughts.

5.1. Content Coding

The recording was transcribed and processed for emergent categories and split into 30 s segments. These classes were organized in the taxonomy of themes according to Figure 4.

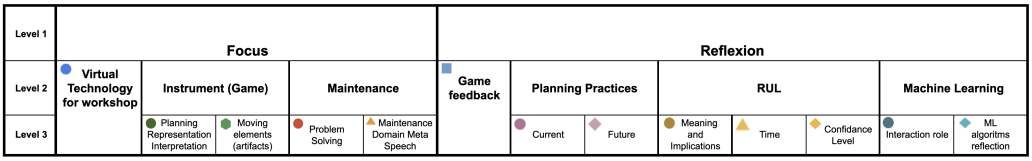

Figure 4. Taxonomy of coded conversation themes.

The speech about *Focus* was found on three immediate themes: on the technological tools used for the workshop; on the planning game exercise, which was further subdivided into the interpretation of how planning is represented, the manipulation of the artefacts; and the conversation related to the game as an instrument. Focus on maintenance was divided into solving the scheduling problem and meta speech related to the maintenance domain.

The speech about *Reflection* of thoughts was divided into four topics: ideas for the improvement of the instrument (game or playful probe); the way the maintenance of the aircraft is being or can be done; the degree of confidence and the meaning/implications that the RUL indicator can have in planning; and finally how machine learning can interact with the human planner or in relation to the operation of ML algorithms. Planning practices is separated into two categories: current practices are currently practised, and future practices are speculated to be implemented according predictive indicators such as RUL. The RUL class is now divided into three subcategories (meaning, time, confidence).

As this is a collaborative work, the discourse of each participant was not grouped by category but by the conversation of the group as a whole. The resulting data are presented in graphs to show how the conversation and the discovery of themes took place throughout the experiment, at what times the themes of RUL and ML were approached, and how the reflections arose, which are important for knowledge development by the participants.

5.2. Conversation Analysis

This experiment lasted 74 min: 23 for Phase 5 and 52 for Phase 6. Before the in-depth analysis, there is an important moment during Phase 5 of the experiment to highlight: 3:00—Problem-setting

The first three minutes were given to the participants to read the maintenance plan and to clear up doubts before moving on to the problem.

The participants began by addressing the problem using meta-speech, suggesting that they were "reading" the problem first and getting the right connection between artefacts and the maintenance language that they were familiar with. For instance, they used the tail aircraft name to refer to specific rows, "there is an issue with the first one, the N100, because it can overlap with the N1003 flight at the beginning". They took about 5 min

between the moment that the problem was placed until they started moving the elements in a very intricate collaboration process, such as analysing and negotiating the movements as if they were learning to play a game of chess. Despite the fact that they assumed different roles and they did not interfere with each other's work, they communicated and collaborated often.

When we look at the focus of the conversation of Phase 5 in Figure 5, we can see that at the beginning, the participants talked about the representation of planning artefacts and have some technical issues related to the technology unfamiliar to them prior to the workshop.

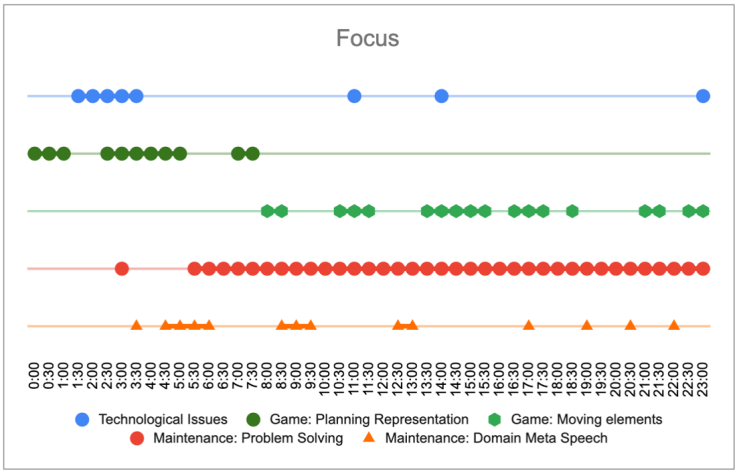

Figure 5. Focus of conversation during Phase 5.

Immediately after the problem was placed, participants started talking about maintenance-related aspects (meta-speech). Only at 5:30 did they change the focus to solve the problem, and only after 8 min did they start to manipulate the artefacts. From this moment onward, the participants did not lose focus on solving the problem until the end of the exercise. This problem resolution was accompanied alternately by moments of artefact manipulation or maintenance-related meta-speech.

The reflection mainly took place during Phase 6, starting at minute 23, immediately after the problem was solved, as can be seen in Figure 6. During this phase, it is important to note that there was a quite intense discussion about maintenance planning practices. The discourse alternates between current practices and speculation on what future practices will look like. The reflective discourse in this phase is divided into three major blocks: Between 23 and 40 min, we found speech oscillating between current and future practices, specifically regarding RUL (e.g., RUL confidence level representation, RUL as a box plot or distribution, maintenance risk and criticality, due date management, state of the fleet, maintenance opportunities, cluster and block management, task and RUL management, and RUL and task management); between 42 and 58 min, the speech was about future practices and mostly focused on ML (e.g., agent suggestions and planner assessment, solution fine tuning, deviation implications, planning constraints and impacts, planner knowledge and strategies, RUL distribution visualisation, impact cost curve visualisation, and operation and maintenance plan integration); between 58 and 70 min, only current practices from a more global management perspective were discussed (maintenance and operational planner communication, operational plan management, maintenance and operational planning times, problem solving, types of maintenance, and flight hour/flight cycle ratios).

Figure 6. Reflections on the conversation during the experiment.

Concerning the introduction of the RUL concept, we could verify that whenever there was a dialogue about time or the confidence interval, it came with a discussion on meanings and implications it may have. This took place mainly in the first block of the mixed discourse between talk of current and future practices.

The discussion about the use of machine learning only started in the last part of the debriefing phase. Due to the participants being involved in the domain of maintenance planning, their initial interest was to discuss and clarify some maintenance concepts and anticipate the possible changes in their daily work. Only after this clarification did they begin to explore how ML can contribute to their work, combining both concepts. Specifically, participants discussed future maintenance practices and the use of ML algorithms to help planners in that task. As shown in Figure 6, both the reflections on interacting with ML algorithms were completely connected to a conversation about future aircraft practices. They are interspersed between the form of interaction and reflection on the functioning of the algorithms and appear at several points in time simultaneously. The third block is exclusively a reflection of current practices. Between the first and the second blocks, a moment of reflection on the game (playful probe) itself takes place, but only for 2 min.

In Figure 7, we can verify the relationship between focus and reflection in maintenance. In Phase 5, the participants were completely focused on the solution of the problem and led a meta-discourse on maintenance, especially in an early stage. During this phase, the participants had only 3 moments of reflection on current practices. During the first block of Phase 6, as mentioned above, it is possible to verify the alternation and balance of the reflective discourse between current and future practices. The predominance of the discourse on future practices takes place during the second quarter and on current practices in the third.

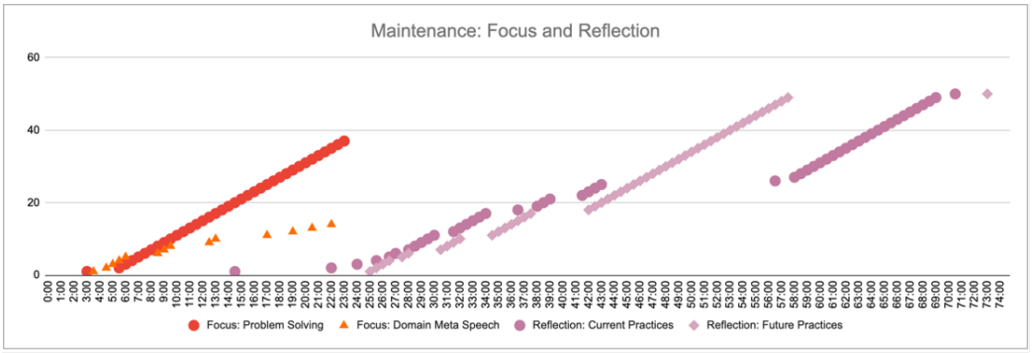

Figure 7. Cumulative number of conversation segments between focus and planning reflection, suggesting 3 different phases.

6. Understandable Interactions

In this section, we describe the conversation about the main interaction themes that emerged during both runs, synthesizing the insights of the evolution of ML interaction towards a CBM planning paradigm.

The discussion is done through analysis of the concrete conversation utterances during the experiment while comparing with feedback obtained from participants' post-session interviews.

Design Insights

In this section, we list and detail the various insights to the iteration that emerged during the conversation with the participants.

1. **Understandable maintenance representation**
   Regarding the experience of interpreting the game elements (flights, blocks, clusters of tasks, and plans) the participants generally found it clear, with P1 adding *"clear and similar to tools already in use"* and P2 saying *"this view is actually quite nice to be able to quickly scan the situation"*. As can be visualized in Figure 5, the participants started to talk about planning representation and then immediately started to move artefacts at minute 8.
   P1, during the exercise, verbalized the possibility of also visualizing other kinds of **maintenance that does not require a hangar**, which is important in a short-term maintenance paradigm such as CBM, concluding *"If it's a really small problem you can do it during a turnaround"*.
2. **Maintenance package management**
   In respect to future developments of maintenance artefacts, P1 was thinking about how to visualise the "benefit you get **combining a cluster and a block**. Let's say, after this 30H (of maintenance), there are 1H of toeing, that means if you combine them, you save an hour, so the box becomes a little bit smaller". P2 was also concerned with this kind of plan optimization, *"we* ***should combine these two****, because it's a kind of waste bringing them back to the hangar twice in two day"*. Both participants were interested in *"open"* and split clusters being used as some sub-clusters, especially if there was a task where some RUL restricts the entire cluster. This should be interesting if there is an update in only one RUL among the possible dozens at any given time, and the best solution is to solve the problem related to this specific RUL and leave the rest of the cluster according to the original schedule, *"in fact,* ***what we necessarily need to move, is not all the work, but part of work.****"* (P2).
3. **Maintenance flexibility and control**
   At some point, P1 considered scheduling two hours of maintenance over the limit, and wondered *"what is the consequences of not making the exact Due Date? what's the consequences of having the component filled before the preventive removal?"* and *"How critical is it if we don't respect a RUL?"*, suggesting that the **planner should have the flexibility to schedule tasks in other time if the return is large enough**. Participants agreed that we should start from the assumption that the planner knows things that cannot be coded in the model, *"the planner might have more data or might have some preferences, some strategies in his head, that make him decide to deviate from the output of ML algorithm"* (P1). Thus, we should assume that s/he can make some changes based on human (tacit) knowledge and turn these into constraints to generate a new solution. This can be done by fixing a particular block or cluster or locking an empty space after some maintenance because s/he *"knows there is a risk that they are working in an area where usually have other findings which they need to attempt too as well"* (P2).
4. **Manual planning**
   Complementing the previous point, P2 said that that **maintainers need some room to schedule clusters because they do not know what kind of corrective tasks they will have in 30 days**: "we don't have the luxury always of having RUL of more than

100 h (…). The problems pop up, let's say, in common flights, so we need to act on that right now (…) to find some spot to fix the next couple of days".

5. **Maintenance time restrictions**
   Participants confirmed the time to fix blocks as part of A-checks is done respecting the time limits presented in this exercise "until like you said, the 20 days to 30 days" (P1). However, "there are also other work as modifications, and those you can foresee months of prompt, let's say if you want to install wifi on the aircraft, what this is not popping up on a short term but that you already know months in advance" and can be scheduled in some check.

6. **Flight and maintenance plan merger**
   Although planners do not visualize flights in maintenance planning, in part because planning flights is currently done in the short term, **they recognized the importance of visualizing flights on the same canvas as maintenance**. After participants played with flights, blocks, and clusters, they suggested improvements to make them more complete, such as including the turnaround and towing time in the flight artefacts. Participants also suggested presented the hours per flight. This information may also be important for cooperation with operational planners. A **task with low probability and very high impact can trigger a discussion** about whether it should be planned, and they should simply accept this schedule if they *"**have a spare aircraft stand bay** or have some buffer in the network"*; otherwise, they will not take this risk, which may lead to cancellation.

7. **The role of automatic planning**
   Participants assumed that there would be some form of automatic planning that would reschedule the entire plan. However, they felt the need to plan only part of the plan. P3 felt the need to have a **button to *"fix the rest"*** once s/he made a few choices. P2 also had the same question, *"How will we be able to **lock some parts not to be changed by AI plan recalculation?"***. P2 recognized that it is difficult to manually optimize a solution, venting *"Wow, this is endless!"*, and concerning plan optimize *"we should combine these two, because it's a kind of waste bringing them back to the hangar twice in two days"*.
   Participants agreed that it might be a good idea for the ML agent to **automatically group tasks into clusters and propose a solution to the planner.** Then s/he must make an assessment and decide what to accept, taking into account that s/he will always be able to adjust the solution that the system has proposed.
   Both participants highlighted a few occasions when the ML agent could be called to present some solution. Participant 2 said that the ML agent should be called *"When a new RUL is introduced, either a change or a new cluster"*. P1 also suggest that *"an initial proposal to cope with a new 'problem' would be nice, indicating the differences the ML propose to make"*. P1 also presented an idea similar to some chess applications to improve the interaction between the user and the ML agent: *"If you select a block, perhaps see the options of what you can do with that block, before moving"*.

8. **Discretionary balance between control and autonomy**
   Participants also expected the tool be useful to generate a solution that not only respects the restrictions but also allows **limiting the search space to a certain period of time or to some selected aircraft**. However, it should show the planner the impact of this limitation. *"For example I gave an 8 h (slack) after maintenance just because sometimes there is an issue, but s/he sees in the planning that it has a quite lot of impact"* (P2). P1 agreed that the planner should be able to get some **kind of score, or even better, the cost of making changes,** *"because maybe there are some biases in behavior or maybe (the planner) is used to do in a certain way."* Further, it must be feasible that this actually helps to achieve better solutions *"not just in time reduction, but also in optimality"*.

9. **Maintenance RUL confidence level**
   The run participants found it easy and clear to understand what needed to be done.

However, when confronted with the RUL confidence level, they found it not easy to interpret, and they considered the RUL as a fixed due date. P2 said it "was quite tricky estimate what risk you took when you interpreted the RUL", while P1 said the representation of RUL required some mental effort to visualize: it "was a bit challenging to determine the due dates for the tasks, it required some mental efforts". P1 added during the exercise, **"the difference between 95 and 99 in my head is not playing a role"**. Despite the difficulty in seeing the impact of the confidence level during the exercise, they made an effort to understand the impact of the confidence level; e.g., P1 said ***"I won't to risk, because 90% is quite high"***.

10. **Maintenance RUL visualization**
Participants suggested automatically visualizing the RUL on the timeline, and P1 also suggested it **would be good to *"visualize operation impact"* such as costs, availability, and the maintenance components**, asking P2 *"But it could actually depend on what's these 65 h based on, right? What of kind components we are talking about?"*. During the exercise, P2 suggested an RUL of 60 h with a confidence level of 90%, "it would be nice if we could see (...) 65+−6 h, than you kind have an idea of how close the edge you are", and when asked if a boxplot could fit, P1 answered *"Yeh, I'm thinking out aloud now, but perhaps instead a square box, it could be a kind of distribution"*. At the end of the exercise, P1 took a co-constructive move and started using the collaboration tool to make some design proposals. S/he started to draw **how this kind of distribution could be**, as shown in Figure 8, **a visual analogy based on how the arrival time is modelled but in this case as a view of the risk.**

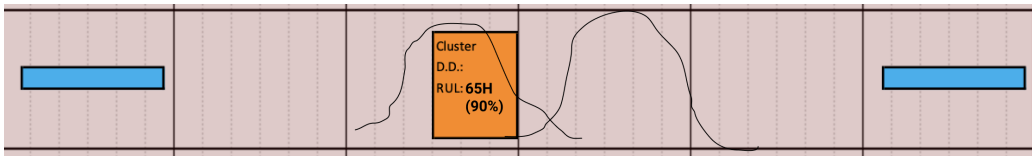

Figure 8. Proposed remaining useful life distribution visualization by Participant 1.

A participants added another curve, shown in Figure 9, and said that it was something that s/he is not used to, that it was just his/her idea based on aircraft management with regards to a future CBM scenario.

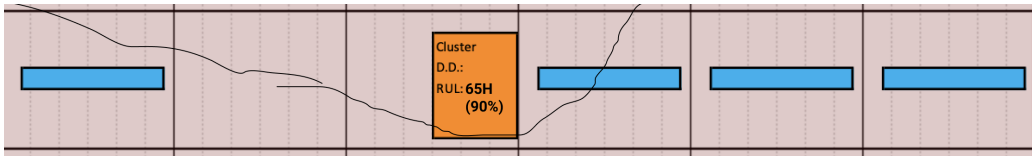

Figure 9. Proposed impact curve by Participant 1.

This should be something related to the **impact cost,** *"so if you do this task now, it will cost you something because it will be based on the RUL (...)* ***if you do it too early it's got a cost because you are wasting the RUL, but if you do it too late it's gonna cost you because it's incurring a delay, cancellation, or high repair times.*** *But there is no optimal here, and there is something that you can play with"*, referring to the possibility of adjusting the best time to schedule some cluster and getting the respective impact of this move.

11. **CBM maintenance indicators**
Participants asked about if they have the **data needed to possibly turn the impact curve**; the participant added *"we know the delay cost, we know the cancellation cost approximately, we know very much escalate repair cost is, and we know about how much RUL cost approximately, what is a bit more difficult it's the cost of preventive repair"*. P2 presented his/her vision: *"**we should have a kind of class of component or class of consequences**, and depending on that class, it must not run the risk, or it can run the*

risk of exhausting the RUL". P1 agreed, *"the decision on whether to schedule something, should not be just dependent on the description of the task but should be also dependent of the maintenance opportunities and the state of the fleet"*, and *"take in consideration the probability that's something might fill with the large or small impact"*.

7. Conclusions

This work showed how the playful probe exercise materialized in a digital paper prototype and enabled an exploratory environment in which researchers and domain experts were able to explore diverse aspects of adoption of ML in the coming practice of CBM in airline maintenance. By focusing on playing with playful artefacts to solve a concrete problem, participants could reflect on changes to their domain and open a speculative and productive dialogue on how CBM maintenance could be designed, as evidenced by content analysis over action and speech during the exercise and presented as understandable interactions. Through the use of playful probes, we were able to raise questions about interacting with an ML planning agent and dealing with RUL estimates.

The way probes are built has a great impact on the workshop and the reflections obtained. Sections 3 and 4 describe the detailed process of how complementary participatory design methods were found by playful probing, and how probes were carefully studied and designed to offer the correct "bubble" experience [9] to participants, which allowed us to originate useful reflections for speculative study of interactions.

This study showed that playful probes, even in a non-dynamic environment such as a paper prototype exercise, can serve as a valuable tool to direct the dialogue to relevant aspects of new interactions as yet to be developed and addressing resources, knowledge, and meaning aspects for that interaction, as evidenced in the analysis of participants' discourse. Playful probes can enable this exploration in a cooperative way, as suggested in the workshops by the participants themselves. This exercise allowed the participants to put themselves in a safe and relaxed environment to play and learn collaboratively how to deal with a high-risk problem.

Author Contributions: Conceptualisation, J.R. and L.R.; investigation, J.R.; methodology, J.R.; supervision, L.R.; validation, L.R.; writing—original draft, J.R.; writing—review and editing, L.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon 2020 research and innovation program under the REMAP project, grant number 769288 and funded by the FCT—Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit—project code UIDP/00326/2020. The first author is also funded by the FCT—Foundation for Science and Technology, under the grant 2022.11131.BD.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CBM	Condition-Based Maintenance
ML	Machine Learning
AI	Artificial Intelligence
AMP	Aircraft Maintenance Planning
RUL	Remaining Useful Life

References

1. Bødker, S.; Kyng, M. Participatory design that matters—Facing the big issues. *ACM Trans. Comput.-Hum. Interact.* **2018**, *25*, 1–31. [CrossRef]
2. Mattelmaki, T.; Korkeakoulu, T. *Design Probes*; University of Art and Design: Helsinki, Finland, 2008.
3. Gaver, W.W.; Boucher, A.; Pennington, S.; Walker, B. Cultural probes and the value of uncertainty. *Interactions* **2004**, *11*, 53. [CrossRef]

4. Celikoglu, O.M.; Ogut, S.T.; Krippendorff, K. How Do User Stories Inspire Design? A Study of Cultural Probes. *Des. Issues* **2017**, *33*, 84–98. [CrossRef]
5. Lange-Nielsen, F.; Lafont, X.V.; Cassar, B.; Khaled, R. Involving players earlier in the game design process using cultural probes. In Proceedings of the 4th International Conference on Fun and Games-FnG '12, Toulouse, France, 4–6 September 2012; ACM Press: New York, NY, USA, 2012; pp. 45–54. [CrossRef]
6. Hutchinson, H.; Hansen, H.; Roussel, N.; Eiderbäck, B.; Mackay, W.; Westerlund, B.; Bederson, B.B.; Druin, A.; Plaisant, C.; Beaudouin-Lafon, M.; et al. Technology probes. In Proceedings of the Human factors in Computing Systems-CHI '03, Ft. Lauderdale, FL, USA, 5–10 April 2003; ACM Press: Ft. Lauderdale, FL, USA, 2003; p. 17. [CrossRef]
7. Vasconcelos, A.; Silva, P.A.; Caseiro, J.; Nunes, F.; Teixeira, L.F. Designing tablet-based games for seniors: The example of CogniPlay, a cognitive gaming platform. In Proceedings of the Fun and Games '12: International Conference on Fun and Games, Toulouse, France, 4–6 September 2012; Volume 3, pp. 1–10. [CrossRef]
8. Wallace, J.; McCarthy, J.; Wright, P.C.; Olivier, P. Making design probes work. In Proceedings of the Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 3441–3450. [CrossRef]
9. Huizinga, J. *Homo Ludens: A Study of the Play-Element in Culture*; Angelico Press: Brooklyn, NY, USA, 2016.
10. Gaver, B.; Dunne, T.; Pacenti, E. Design: Cultural probes. *Interactions* **1999**, *6*, 21–29. [CrossRef]
11. Sahay, A. An overview of aircraft maintenance. In *Leveraging Information Technology for Optimal Aircraft Maintenance, Repair and Overhaul (MRO)*; Elsevier: Amsterdam, The Netherlands, 2012; pp. 1–230. [CrossRef]
12. Knowles, M.; Baglee, D.; Wermter, S. Reinforcement learning for scheduling of maintenance. In *Res. and Dev. in Intelligent Syst. XXVII: Incorporating Applications and Innovations in Intel. Sys. XVIII-AI 2010, 30th SGAI Int. Conf. on Innovative Techniques and Applications of Artificial Intel.*; Springer: London, UK, 2011; pp. 409–422. [CrossRef]
13. Andrade, P.; Silva, C.; Ribeiro, B.; Santos, B.F. Aircraft maintenance check scheduling using reinforcement learning. *Aerospace* **2021**, *8*, 113. [CrossRef]
14. Bernhaupt, R.; Weiss, A.; Obrist, M.; Tscheligi, M. Playful probing: Making probing more fun. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 4662 LNCS*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 606–619. [CrossRef]
15. Sjovoll, V.; Gulden, T. Play probes-As a productive space and source for information. In Proceedings of the 18th International Conference on Engineering and Product Design Education: Design Education: Collaboration and Cross-Disciplinarity, E and PDE 2016, Aalborg, Denmark, 8–9 September 2016; Number September; The Design Society: Copenhagen, Denmark; Institution of Engineering Designers: Glasgow, UK, 2016; pp. 342–347.
16. Kjeldskov, J.; Gibbs, M.; Vetere, F.; Howard, S.; Pedell, S.; Mecoles, K.; Bunyan, M. Using Cultural Probes to Explore Mediated Intimacy. *Australas. J. Inf. Syst.* **2004**, *11*, 102–115. [CrossRef]
17. Moser, C.; Fuchsberger, V.; Tscheligi, M. Using probes to create child personas for games. In Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology-ACE '11, Lisbon, Portugal, 8–11 November 2011; ACM Press: New York, NY, USA, 2011; p. 1. [CrossRef]
18. Klabbers, J.H.G. *The Magic Circle: Principles of Gaming & Simulation*; Modeling and simulations for learning and instruction; Sense Publishers: Rotterdam, The Netherlands, 2006.
19. Ribeiro, J.; Roque, L. Playfully probing practice-automation dialectics in designing new ML-tools. In Proceedings of the VideoJogos 2020: 12th International Conference on Videogame Sciences and Arts, Mirandela, Portugal, 26–28 November 2020; pp. 1–9.
20. Ribeiro, J.; Andrade, P.; Carvalho, M.; Silva, C.; Ribeiro, B. Playful Probes for Design Interaction with Machine Learning: A Tool for Aircraft Condition-Based Maintenance Planning and Visualisation. *Mathematics* **2022**, *10*, 1604. [CrossRef]
21. Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K.; et al. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems-CHI '19, Glasgow, UK, 4–9 May 2019; pp. 1–13. [CrossRef]
22. Holbrook, J. Human-Centered Machine Learning. 2017. Available online: https://medium.com/google-design/human-centered-machine-learning-a770d10562cd (accessed on 16 April 2020).
23. Guzdial, M.; Liao, N.; Chen, J.; Chen, S.Y.; Shah, S.; Shah, V.; Reno, J.; Smith, G.; Riedl, M.O. Friend, collaborator, student, manager: How design of an AI-driven game level editor affects creators. In Proceedings of the Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–13.
24. Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B.Y.; Kankanhalli, M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; ACM: Montreal, QC, Canada, 2018; pp. 1–18. [CrossRef]
25. Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. Designing Theory-Driven User-Centric Explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems-CHI '19, Glasgow, UK, 4–9 May 2019; ACM Press: Glasgow, UK, 2019; pp. 1–15. [CrossRef]
26. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **2021**, *10*, 593. [CrossRef]
27. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **2021**, *23*, 18. [CrossRef] [PubMed]

28. Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J.M.F.; Eckersley, P. Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; ACM: Barcelona, Spain, 2020; pp. 648–657. [CrossRef]
29. Bødker, S.; Roque, L.; Larsen-Ledet, I.; Thomas, V. Taming a Run-Away Object: How to Maintain and Extend Human Control in Human-Computer Interaction? In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, 21–26 March 2018; pp. 1–6.
30. Lukosch, H.K.; Bekebrede, G.; Kurapati, S.; Lukosch, S.G. A Scientific Foundation of Simulation Games for the Analysis and Design of Complex Systems. *Simul. Gaming* **2018**, *49*, 279–314. [CrossRef] [PubMed]
31. Vaishnavi, V.K.; Purao, S. (Eds.) Design Science Research in Information Systems. In Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, DESRIST 2009, Philadelphia, PA, USA, 7–8 May 2009.