



Human-centric algebraic machine learning

Interaction Paradigm Methodology - Final

Deliverable D4.6

WP4 - Human AML Interaction



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952091.

Consortium



ALGEBRAIC AI



Champalimaud
Foundation



Project

Title: ALMA: Human Centric Algebraic Machine Learning
 Acronym: ALMA
 Coordinator: Proyectos y Sistemas de Mantenimiento SL (eProsima)
 Grant agreement: 952091
 Call: H2020-EIC-FETPROACT-2019
 Program: Horizon 2020
 Start: 01 September 2020
 Duration: 54 months
 Website: <https://alma-ai.eu>
 E-mail: alma@eprosima.com
 Consortium: **Proyectos y Sistemas de Mantenimiento SL (eProsima) (EPROS)**, Spain
Champalimaud Foundation (CF), Portugal
ALGEBRAIC AI (ALG), Spain
German Research Center for Artificial Intelligence (DFKI), Germany
Inria (Inria), France
TU Kaiserslautern (TUK), Germany
FIWARE Foundation e.V. (FF), Germany
Technical Research Centre of Finland (VTT), Finland
Carlos III University (UC3M), Spain

Deliverable

Number: **D4.6**
 Title: **Interaction Paradigm Methodology - Final**
 Month: 54
 Work Package: WP4 - Human AML Interaction
 Work Package leader: Inria
 Deliverable leader: National Institute for Research in Digital Science and Technology (Inria)
 Deliverable type: R: Document, report
 Dissemination level: Public (PU)
 Date of submission: 2025-02-28
 Version: v2.0
 Status: Finished

Version history

Version	Date	Responsible	Author/Reviewer	Comments
v1.0	2025-01-28	Inria	draft	
v2.0	2025-02-28	Inria	final	

Executive summary

AML Algebraic Machine Learning (AML) is a novel machine learning technique based on algebraic representations of data. Unlike statistical learning, AML algorithms are robust regarding the statistical properties of the data and are parameter-free. AML systems can retain what they learn, including sociocultural ethical practices for human–artificial intelligence systems. The aim of the EU-funded ALMA project is to study the mathematical properties of AML and to leverage them to develop a new generation of interactive human-centric machine learning systems. These systems are intended to minimize bias in order to facilitate trust while better enabling distributed collaborative learning.

This report corresponds to *Deliverable D4.6 - Interaction Paradigm Methodology - Final* of the ALMA project. The goal of this deliverable is to extend current participatory design and generative design methodologies to incorporate the role of an intelligent system in the interaction loop, where agency shared between the human user and the AML system.

Contents

Executive Summary	3
Contents	4
Acronyms	5
1 Introduction	6
2 Participatory Design	7
3 Methods for Intelligent System Design	9
4 <i>Comparative Structured Observation</i>	11
5 Defining <i>Comparative Structured Observation</i>	13
6 Comparing <i>Comparative Structured Observation</i> to Other Methods	15
7 Characteristics of <i>Comparative Structured Observation</i>	19
8 <i>Comparative Structured Observation</i> Criteria	21
9 Applying <i>Comparative Structured Observation</i>	22
9.1 STORYLINES Project	22
9.2 STORYLINES Study	22
9.2.1 Participants:	23
9.2.2 Control room organization:	23
9.2.3 Participatory design workshops:	24
9.3 DESIGNPROMPT Project	26
9.4 Participants	27
9.5 Results	27
10 Discussion	28
11 Conclusion	29
References	30

Acronyms

AML Algebraic Machine Learning.

1 Introduction

The goal of this deliverable (*Deliverable D4.6 - Interaction Paradigm Methodology - Final*) is to explore how to extend current participatory design and generative design methodologies to include an intelligent agent in the interaction loop. The key challenge is how to manage the sharing of agency between human users and the AML system, in such a way that increases the efficacy of both. This requires new research and design methods that emphasize the *interaction* between the human user and the AML system, as opposed to the qualities of the algorithm alone.

(*Deliverable D4.5* discussed the challenges of creating *human-centered AI*, that is, intelligent interactive systems with a measurably positive impact on human users. The key is to create “*human-computer partnerships*” that combine human skills and system capabilities so that humans and intelligent agents perform better together than either individually. This requires not only taking advantage of the best characteristics of each, but also considering how to improve human capabilities over time, thus “upskilling” rather than deskilling or replacing the users of these systems [1]. Unfortunately, that most current AI research is measured in terms of improvements to an algorithm, not on the effect of the intelligent system on human users over time.

This became an issue of concern to society at large with the introduction of CHATGPT, which fundamentally changed the ways in which the general public viewed artificial intelligence in general, and machine learning in particular. Being able to write text prompts, ask questions or sketch lowered the barrier to entry for taking advantage of such systems. While this level of access is very useful for a variety of applications where the goal is to gather inspiration or get suggestions for text or images, it does not address a key underlying problem i.e. that users do not have any way of knowing how the system works or verifying the accuracy of the information they receive.

CHATGPT, *Dall-E* and more recent systems represent major technical breakthroughs, they also carry many embedded assumptions about how to interact with them. Asking people to converse with a simulated person appeals to our very human tendency to anthropomorphize the objects and animals we interact with. However, this can lead to severe problems when the intelligent agent is wrong or “hallucinates” [2]. The term “hallucination” was introduced to describe any incorrect or misleading results that AI models generate. These errors can be due to incorrect model assumptions, insufficient training data, or biases in the model’s training data. Unfortunately, hallucinations are altogether too common and are difficult for even expert users to identify.

This leads to a fundamental challenge for designers of intelligent systems: How can we create intelligent systems that users, whether experts in the domain or not, can understand? How can users identify a system’s capabilities and potentials for error, so they can use them effectively? How can users understand what the system can do for them?

After first describing the history and value system underlying participatory design, this deliverable describes how the *Comparative Structured Observation* method can help designers assess the role of an intelligent system in the interaction loop, where agency is revealed and shared by the human user and the AML system.

2 Participatory Design

Susanne Bødker [3] identifies three “waves” of human-computer interaction (HCI) research. The first wave of HCI focused on building systems that were easy to use, with quality judged from an engineering perspective. Researchers and developers ran controlled experiments with quantitative measures, not only of the machine’s performance, but also of the human user’s performance. Traditional measures included speed of execution and rate of errors, as well as questionnaires as to how well the user’s liked the system. This style of HCI research made sense in an era when computers were large mainframes focused on complex computational tasks in government or corporate settings.

The second wave of HCI shifts the focus from the computer to the human user, who is viewed as an “information processor” and “decision maker”, using psychology and cognitive science as its theoretical underpinnings. Here, the goal was to model user behavior and create designs informed by those models. This era was accompanied by massive changes in computing hardware and software, where white collar workers could operate their own “personal” computers at work. New conferences, such as “Computer Supported Cooperative Work” began to consider how groups of people collaborate with computers, although the emphasis remained on work, with input from new research disciplines, such as sociology.

The third wave of HCI was accompanied by the advent of personal computing, and the widespread use of a wide variety of computing devices, including smart phones and watches, for both work and personal use. The research focus changed to incorporate the social-cultural context of computing, which required contributions from still more research disciplines as well as more active contributions from all aspects of design.

Each wave responded to the significant changes in hardware, software and network capabilities, but also in the purposes and settings in which they were used. These fundamental changes in computing has required corresponding changes in HCI research and design methodologies, which shifted from studying the “human factors” related to making more efficient, “user friendly” systems to considering how to integrate computing into every aspect of our lives.

The earliest HCI research supported traditional software development approaches from the 1970s and 1980s. During this period, software engineers imitated the hardware engineering process, which demanded a complete specification of the final system up front, followed by a systematic development process that produced the final result. The corresponding software development process, often referred to the “water-fall approach”, involved beginning with a detailed specification of “user requirements” usually phrased in terms of user tasks that are related to specific computer functions. After these functions have been implemented into the final system, the developers may run usability studies, with quantitative measures of performance and/or questionnaires to learn what users liked or disliked about the system.

Participatory design was developed in the Scandinavian countries at the same time, but took a completely different approach that actively involved users as key stakeholders throughout the design process. According to Spinuzzi [4], participatory design is a research paradigm with “its own highly articulated methodological orientation, methods, and techniques ... [with] a commitment to sustained, methodical investigation according to grounded methodological principles.” HCI researchers who were influenced by this approach focused on understanding the users’ needs in the real-world context of use.

They emphasize qualitative methods for exploring design possibilities with users to both co-design new design ideas and test them with users. Quantitative methods, in the form of controlled experiments, tend to be used primarily in the evaluation phase of the design process. Another key characteristic of participatory design is to treat users as important stakeholders in the design process, from problem identification and developing research questions to disseminating the final results. Note that User Experience (UX) design practitioners use a related term – [co-design] – to refer to ensuring that users are consulted during the design process, although not always as stakeholders.

Both participatory design and co-design fit well with *Agile*[5] programming, which emphasizes the ability to create and respond to change. Unlike the traditional “waterfall” software development methods, the goal is to be receptive to user needs and to adjust as circumstances change, rather than attempting to define all aspects of the system from the outset.

Clement and van den Bessar clarify the differences between participatory design and more conventional research approaches:

Unlike conventional research, which is directed primarily at producing results of interest to those beyond the immediate research site, an essential goal of action research is to achieve practical or political improvements in the participants’ lives, e.g., less routine work, greater autonomy, more effective tools. The researcher becomes directly involved in the ongoing work and feeds results back to the participants.” (Clement and van den Bessar, 1993, p.33.)

Spinuzzi [4] identifies the following criteria for participatory design, drawn from the Scandinavian tradition:

1. *Improve the Worker’s Quality of Life* by increasing the worker’s control over their own work organization, tools and processes, and improving their ability to perform tasks with ease. This requires a critical reflection on tacit knowledge, invisible work and unstated individual and organizational goals. Goals include reaching agreement through interviews, focus groups, workshops, organizational games, prototyping sessions, and other techniques and co-determining the project goals between researchers and users.
2. *Collaborative development* to explore and articulate tacit knowledge and invisible practices, where users participate in both data collection and analysis. Researchers ensure that user’s are represented and provide routes for ensuring that their concerns are addressed.
3. *Iterative process* to ensure a sustained reflection on the designed artifact, as each iteration reveals new invisible work practices and tacit knowledge. This requires co-design activities throughout the design process, with a flexible approach for handling change.

What is clear from the above is that participatory design has a deep political foundation with clearly articulated humanistic values. The goal is explicitly to empower workers by actively integrating them as key stakeholders into the design process. While the term is used more broadly for a wider variety of design activities, it is useful to remember the reasons for why certain aspects of participatory design were chosen.

From an interaction design perspective, the goal is to create intelligent interactive systems that truly meet the needs of their users. We argue that this requires an approach that involves active participation by users, rather than the more common approach of developing the system first and then testing it on users. Valuing human perspectives and explicitly seeking to incorporate their tacit knowledge about the domain and the context of use can result in significantly better AI, where human users take advantage of uniquely human capabilities and benefit from the resources offered by AI, rather than simply becoming another cog in the machine.

3 Methods for Intelligent System Design

Our own work on participatory design methods has evolved over the past few decades ([6, 7, 8] and has included a strong emphasis on the value of multi-disciplinary mixed methods [9, 10] While we retain the emphasis on working closely with users, our primary goal is to generate truly novel designs that demonstrate new principles for how to interact effectively with different forms of computers. We also explore a much wider range of users and use settings, since we are interested in creative and personal uses of computing, in addition to work settings.

However, we have also developed what we call “generative theories of interaction” [11], which are inspired by existing theories of human cognition and behavior and applied to interaction design. These theories include concepts specific to Human-Computer Interaction (HCI) that can be transformed into actionable principles for designing interactive systems. For example, we explore how to transform theories from other disciplines such as co-adaptation from evolutionary biology [12] or technical reasoning from ecological psychology [13, 14] into actionable principles for design.

Applying a generative theory of interaction involves three basic steps:

1. **analyze** the system according to the design principles.
2. **critique** specific elements of the system, and finally
3. **construct** a new design that more fully integrates the above design principles.

These generative theories offer both researchers and User Experience (UX) designers the opportunity to design truly novel forms of interaction that are both simpler to use and provide greater power of expression. However, we argue that the advent of generative AI raises the need for additional generative theories, with corresponding research methods, that allow us to integrate machine learning into interactive systems in ways that truly meet users needs.

How does AI change participatory design?

The basic participatory design approach and value system does not change: we must still take the user’s perspective into account and include them throughout all phases of the design process. However, a key problem for HCI researcher and UX designers is that AI makes it much more difficult to predict what the machine learning algorithm will produce. It is harder to clearly identify edge cases and address the potential harm of unexpected hallucinations. One of the hallmarks of AI systems is that users tend to over-estimate their reliability, which can lead to serious problems. Designers need to include this potential for the system to generate errors, which may be subtle and hard to detect, and to mitigate the dangers of having users over-rely on systems.

Another key design challenge is how to avoid transforming users from active agents who are in control of the interaction, into more passive recipients of suggestions from the AI, effectively becoming error checkers rather than authors. From a human perspective, poorly designed interaction with an intelligent system can severely reduce the quality of the user’s job. The cost of generating new information is replaced by the cost of checking for errors.

How can we measure the effect of incorporating AI?

Many of the existing quantitative methods that have been developed by the HCI research community or borrowed from other disciplines can be used effectively to evaluate an intelligent system once it has been implemented. Designers can also use standard qualitative methods for gathering information about users and generating novel ideas. However, the field lacks a clearly identified, rigorous qualitative method for assessing a design concept or choosing among different design variants, because at this point, the designer is not interested in measuring performance – *getting the design right* – but rather in an earlier question, which is *getting the right design* [15] The next section describes our work on *Comparative Structured*

Observation, a novel qualitative mixed method that provides a structured approach for gathering design insights from users who have experienced grounded, contextually relevant tasks with different design variants and asking for their reflections on the complex sets of trade-offs that obtain.

4 *Comparative Structured Observation*

A key role of Human-Computer Interaction (HCI) research is to generate novel concepts that inform the design of interactive systems. HCI researchers can *generate* new design insights by drawing upon a variety of well-defined qualitative methods in the early phases of a design-based research project. For example, naturalistic observation helps researchers understand user needs in field settings and produce implications for design; whereas participatory design involves target users and researchers working together to explore new design concepts. However, *assessing* a nascent design concept or verifying a design direction remains challenging, especially since the goal is not yet to validate a final design, but rather to obtain insights from users that help refine the concept.

A common perception among HCI researchers is that evaluating a new system design requires running a controlled experiment that compares it to other designs according to pre-defined quantitative measures by establishing a causal connection between the proposed design and dependent, quantitative measures. However, such experiments are only appropriate for evaluating clearly articulated, testable hypotheses, such as demonstrating improved performance with a specified interaction technique. They are far less relevant when the researcher's goal is to assess the user's subjective reactions to potential design variants or to choose among alternative design directions. A problem arises when a research project is advanced enough to have an initial design concept, but a quantitative assessment is either too early or altogether inappropriate.

This assumption that HCI publications require quantitative experiments, even if such experiments do not fully assess or advance the design, may lead authors to avoid qualitative study designs, even when quantitative assessments are either too early or altogether inappropriate. Greenberg and Buxton [15] argue that this emphasis has a cost, since it encourages HCI researchers to focus on *getting the design right*, i.e. selecting a design early and perfecting it, rather than *getting the right design*, i.e. generating and comparing multiple design alternatives to explore their strengths and weaknesses. Prominent HCI researchers have gone further, speculating that HCI's over-emphasis on quantitative usability evaluations may be linked to the low adoption of published interaction design research outside of the field [15, 16, 17]. Somewhat informal qualitative usability studies exist, such as industry studies that gather users' reactions and opinions about a new design, however, they are often discounted by reviewers as insufficiently "scientific".

Although designers in industry often run qualitative usability studies to gather users' reactions to and opinions about a new design, such studies are rarely submitted to HCI research venues because they are likely to be discounted by academic reviewers as insufficiently "scientific". This leaves researchers interested in *getting the right design* with a dilemma: Is it better to run an informal qualitative study that produces "fuzzy" claims about how much users "like" the new design or to run an inadequately controlled experiment with hypotheses chosen more to demonstrate statistical significance than for their relevance to assessing the design concept? Should the authors add an experiment to appease reviewers even if the primary contribution stems from qualitative insights?

Some authors highlight qualitative insights as their key research contribution, but add an experiment to appease reviewers. Worse, some reviewers may find themselves wanting to accept an article, but end up rejecting it because the inadequacy of the experiment detracts from the work.

We argue that the HCI research community needs a clearly articulated qualitative method that occurs after exploratory research has resulted in a promising design direction but before the fully implemented design can be validated. The goal here is to produce a rigorous, qualitative assessment of the new concept, where users can experience, compare and reflect on design variants, or compare the proposed design to the status quo. Comparison is central to Structured Observation. It enables not only multiple design variants to be assessed (to increase likelihood of getting the right design) but it can mitigate nuisance variables, including the Hawthorn Effect.

We have developed a novel mixed method called *Comparative Structured Observation* [10] which is designed for use in the middle phases of the design process, when the designer is trying to assess a particular design direction or to compare the trade-offs across multiple design methods. Researchers observe users and ask them to compare and reflect on each variant. They then apply a set of criteria for creating a *Comparative Structured Observation* study and illustrate variations of the method by analyzing four published studies. We also examine five additional studies (two “near” and three “far”) to clarify the boundary between what should and should not be considered a *Comparative Structured Observation*. We discuss the benefits and limitations of the method and argue that gathering comparative reflections about design variants can help researchers assess and advance their design concepts.

The method takes advantage of the structure of controlled experiments to generate and compare grounded reflections from users about design variants. Rather than testing hypotheses about causal relationships between factors, the goal is to advance the design concept. The *Comparative Structured Observation* method involves systematic observation of participants’ behavior under ecologically relevant conditions and encourages both participants and HCI researchers to compare and reflect upon the similarities and differences across multiple design variants.

5 Defining *Comparative Structured Observation*

The primary goal of *Comparative Structured Observation* is to **capture insights from participants’ reflections about selected design variants in order to generate implications for design**, either by identifying the aspects of the design that do or do not work well and why or by providing direction for how to improve the design. **Comparison is essential** — both participants and researchers must be able to compare and reflect on the advantages and disadvantages of each variant relative to its context of use. Because *Comparative Structured Observation* often yields knowledge about how tasks moderate the success of different design choices, sometimes characterized as *What design works when? How? and Why?*, design implications should be richly contextualized.

We define *Comparative Structured Observation* as:

an interventionist, qualitative method for assessing and advancing a design concept where researchers observe participants as they compare and reflect deeply upon their experiences with selected design variants, exposure to which is structured according to experimental design principles.

We chose the name *Comparative Structured Observation* because it evokes key desired characteristics of the method: It emphasizes qualitative observation of how study participants experience and compare design variants, with structured tasks or events that are usually imposed by the researcher. Although sociologists use the term *Structured Observation* informally, mostly to remind researchers that creating a structured observation checklist will facilitate subsequent data analysis. We feel that the longer term *Comparative Structured Observation* both captures what we mean and will not be confused with prior and informal uses of the shorter term *Structured Observation*.

Our definition of *Comparative Structured Observation* is explicitly interventionist, since the researcher imposes design variants for the study participants to experience, either by generating tasks for them to perform or ensuring the presence of naturally occurring events.

Comparative highlights the centrality of comparison, which is key to the design of controlled, quantitative experiments [18]. Here, we obtain similar benefits with qualitative studies, since exposing participants to two or more design variants encourages them to reflect on and compare the strengths and limitations of each. Participants explicitly compare their experiences with ecologically grounded tasks which provides a rich foundation for critiquing the proposed design elements, with a correspondingly richer, deeper feedback.

When participants “think aloud” [19, 20], they think more deeply about their tasks. Asking participants to compare makes it easier for them to critique design elements with respect to their own experiences and use context, which is essential for achieving richer, deeper feedback. Comparison also helps researchers assess the impact of design decisions more clearly, especially those that vary across the design variants.

Structured refers to the structuring of study participants’ activities and the ordering of their exposure to different design variants, e.g. through counter-balancing. The researcher selects specific tasks or events to facilitate both the participant’s and the researcher’s ability to compare and reflect on the variants. The tasks should be grounded in real-world context, in other words, be as ecologically valid as possible. The researcher takes advantage of the principles of experimental design to control the presentation of conditions (the design variants) and tasks as well as to explicitly account for various “threats to validity” [18, 21], such as primacy and recency effects, learning and fatigue effects, and confirmation bias [22]. Note that borrowing the structure of controlled or quasi-experiments [18, 21] does not imply that we also borrow the corresponding emphasis on quantitative measures, nor the expectation that the studies should be designed to test operationalized hypotheses or establish causal relationships.

Observation highlights researcher’s observation and recording of qualitative aspects of the users’ behavior. However, such observations need not be co-present or even live. Researchers could set up a remote

observation via a video link, log participants' interactions, or ask participants to use a think aloud protocol [23] as they perform tasks that are recorded and assessed later. Researchers and participants must interact with each other, *Comparative Structured Observation* requires interaction between the researcher and the participant, usually in the form of a post-hoc interview, in order to gather detailed reflections about participants' in-context experiences with the system of interest. Note that interviews alone are insufficient to be considered "observation"—in all cases, researchers need an independent source of data that captures each participant's behavior with respect to the design. This combination of observation and interviews aligns well with the mixed-method framework advocated by Johnson & Onwuegbuzie [22], where interviews conducted during or after observing how participants have performed tasks or experienced events add to the richness of the qualitative results that can be drawn from the study.

We chose the term *Comparative* because of the centrality of comparison to the method. Participants must be exposed to two or more design variants, so they can reflect on and compare the strengths and limitations of each. Asking participants to compare opens the possibility for them to critique design elements, which is essential for achieving deeper more fulsome feedback. Comparison is also critical for the researchers to see more clearly the impact of design decisions, especially those that vary across the design variants.

We chose the term *Structured* to refer to the structuring of study participants' activities and their exposure to different design variants (or conditions) in the study. The researcher selects specific tasks or events to facilitate both the participant's and the researcher's ability to compare and reflect on the variants. The researcher can take advantage of the principles of experiment design to control the presentation of conditions and tasks, as well as explicitly account for various "threats to validity" [18, 21], such as primacy and recency effects, learning and fatigue effects, and confirmation bias [22]. Note that borrowing the structure of controlled or quasi-experiments [18, 21] does not imply that we also borrow the corresponding emphasis on quantitative measures, nor the expectation that the studies should be designed to test operationalized hypotheses or establish causal relationships.

We use the term *Observation* to ensure that the researcher observes and records qualitative aspects of the participants' behavior. However, such observations need not be co-present or even live. Researchers could set up a remote observation via a video link, log participants' interactions, or ask participants to use a think aloud protocol [23] as they perform tasks that are recorded and assessed later. However, *Comparative Structured Observation* does require interaction between the researcher and the participant, usually in the form of a post-hoc interview, in order to gather detailed reflections about participants' in-context experiences with the system of interest. Note that interviews alone are insufficient to be considered "observation"—in all cases, researchers require an independent source of data that captures each participant's behavior with respect to the design. This combination of observation and interviews aligns well with the mixed-method framework advocated by Johnson & Onwuegbuzie [22], where interviews conducted during or after observing how participants have performed tasks or experienced events add to the richness of the qualitative results that can be drawn from the study.

6 Comparing *Comparative Structured Observation* to Other Methods

As a highly multidisciplinary field, HCI draws research methods and analysis techniques from a variety of research disciplines, not only from the social and natural sciences, but also from design and engineering [9, 11]. The disparity across research goals and assumptions can create confusion and conflicts as HCI researchers with different disciplinary backgrounds attempt to justify their research approaches. In a special issue of the journal *Human Computer Interaction*, Gray & Salzman [24] analyze and critique five highly cited HCI articles that used experiments to compare usability evaluation methods (UEMS). Although these experiments had been used to justify advice to practitioners, they found that “small problems in the way these experiments were designed and conducted call into serious question what we thought we knew regarding the efficacy of various UEMs”. They highlight the problem of trying to determine causal relationships between complex, multivariate alternatives when standard quantitative methods, and cite Cook & Campbell’s [18, 21] seminal work on quasi-experimental designs for conducting maximally rigorous field experiments, when researchers seek ecological validity but cannot fully control the conditions experienced by study participants.

Cook & Campbell first summarize the rules and assumptions regarding traditional laboratory experiments, which they treat as the gold standard for establishing cause-and-effect relationships. They explain how the experimenter structures tasks assigned to participants, either by fixing them, randomizing them, or varying them systematically, such as when using a latin square to counterbalance the presentation of conditions to mitigate unwanted order effects including learning and fatigue. They then proceed to define quasi-experiments, which seek to structure the study conditions so as to optimally address what they call “threats to validity”.

Runkel & McGrath [25, 26] compare eight research methods, ranging from formal theory, to controlled experiments, to field studies, and show the choice of any particular method optimizes one kind of result at the expense other results. For example, a controlled lab experiment offers measurement precision at the expense of ecological validity and generality, whereas a field study offers ecological validity at the expense of measurement precision and generality. Because no single method can provide comprehensive findings, Mackay & Fayard [9], advocate a mixed methods, triangulation approach, where researchers choose the combination of methods most appropriate to their research goals, given the current phase of the research project. They also highlight a key distinction between social science goals, which seek better understanding of human behavior, cognition, emotion, movement, etc. and HCI, which seeks to provide implications for the design of interactive systems in light of that understanding.

Both Cook & Campbell and Runkel & McGrath take a positivist stance, which is echoed in Gray & Salzman’s critique of usability studies found in HCI research. Since then, the field has become increasingly open to qualitative research, in an interpretivist framework. To be clear, these discussions are intended for HCI researchers, not UX practitioners, who use the term “usability study” to refer to a wide variety of different kinds of studies, from highly formal and controlled to informal and quick, but usually chosen for pragmatic reasons to assess the current state of the product design.

Johnson & Onwuegbuzie [22] also advocate a mixed methods approach. They compare and contrast the advantages and disadvantages of purely qualitative and quantitative methods in educational research, and advocate using “multiple approaches in answering research questions, rather than restricting or constraining researchers’ choices.” They reject dogmatism and argue that research designs should be “expansive and creative”, and suggest that “researchers take an eclectic approach to method selection and the thinking about and conduct of research.” Perhaps not surprisingly, they represent a field with a strong applied element/textemdasheducational research/textemdashthat seeks to improve educational outcomes. Other applied fields share this goal: management science seeks better methods for managing organizational behavior, and medical research seeks better health outcomes. HCI research has a similar

Research methods	Source Discipline	Project Phase	Research Setting	Primary Research Goal	Inter-vention?	Primary Analysis	Comparison Central?	Controlled Comparison?	Participant Reflection?
Comparative Structured Observation	Psych, Soc Sci	later	any	design implications	yes	qualitative	yes	yes	yes
Naturalistic Observation	Anthro	early	field	understand users	no	qualitative	-	-	-
Contextual Inquiry	HCI	early	field	understand users	yes	qualitative	-	-	yes
Experience Sampling	Soc Sci	all	field	understand users	yes	qualitative	-	-	yes
Diary Study	Soc Sci	all	field	understand users	yes	qualitative	-	-	yes
Speed Dating	HCI, Design	early	lab	design implications	yes	qualitative	yes	-	yes
Cultural Probe	HCI, Design	early	field	idea generation	yes	qualitative	-	-	-
Technology Probe	HCI	later	field	evolve design	yes	qualitative	-	-	yes
Quasi Experiment	Soc Sci	later	field	causal relationships	yes	quantitative	yes	yes	-
Controlled Experiment	Psych	later	lab	causal relationships	yes	quantitative	yes	yes	-
Usability Study	Ergo, HCI	later	any	identify problems	yes	quantitative	-	possible	-

Table 2: Contrast between *Comparative Structured Observation* and other methods common in HCI research, including source discipline; research phase, setting, and goals; intervention and analysis type, and whether the method explicitly requires comparison and reflection.

underlying goal, which is to improve the design of interactive systems, and benefits from a similar, mixed-method approach. Perhaps not surprisingly, the term “structured observation” appears in these more applied research areas, but as previously noted, their use differs from our characterization of *Structured Observation*, which is explicitly interventionist.

Many HCI experiments employ quantitative measures, often of performance, such as speed and error, with standard statistical tests, such as ANOVA (Analysis of Variance), to determine if, for example, one interaction technique is better than another. Researchers sometimes accompany such experiments with questionnaires, e.g. five-point Likert-style questions, to determine participants’ preferences and may also ask more open-ended questions to gain deeper insights into the participants’ perceptions of the system. Experiments work well when the researcher can clearly identify a quantifiable goal, such as improved performance, and can construct a controlled laboratory setting in which to test it.

However, many HCI researchers are not particularly interested in quantitative measures of performance, and instead are interested either in obtaining insights that lead to new design directions, or less formal ways of assessing which aspects of a particular design resonate best with users. These researchers do not have, nor do they seek testable hypotheses. Even so, they need more rigorous ways to structure research studies, beyond simply observing users and asking them if they like it.

Tukey [27, 28] argues that researchers need both exploratory and confirmatory analyses when analyzing experimental data. He highlights the fact that “ideas come from previous exploration more often than from lightning strokes” and that “finding the question is often more important than finding the answer”. He views exploratory data analysis as “an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques”, and which he contrasts with confirmatory data analysis, which, although easier to teach and computerize, requires the most careful planning.

Table 2 compares and contrasts *Structured Observation* to ten empirical methods with well-defined characteristics commonly used by HCI researchers, including some borrowed from psychology and the social sciences, and some created specifically by HCI researchers. We do not include general methodological approaches, such as participatory design [29], action research [30] and activity theory [31], since they encompass a broad range of specific methods and are less amenable to the precise characterization given in the table. Because many empirical HCI studies can be viewed as “user studies”, we omitted them as too general and poorly defined. Similarly we omitted “field studies” as being an overly general term for user studies conducted in the field.

Naturalistic observation [32] is used extensively in anthropology and other social sciences to collect field data, without intervention or other manipulation by the researcher. The primary goal is to understand naturally occurring human behavior, primarily by analyzing qualitative data, with no explicit comparisons or reflections by participants. HCI researchers often use naturalistic observation to discover user needs and prepare for more highly structured design methods.

Contextual Inquiry [33] is an anthropology-inspired HCI method where researchers interview people in a field setting, usually a corporation, to identify their needs with respect to a new or existing technology. The primary goal is to better understand user needs in context, with intervention in the form of interview questions. Although comparison is not required, study participants are sometimes asked to reflect on their experiences with respect to other systems.

Experience sampling [34] and *Diary studies* [35] are borrowed from the social sciences. They ask participants to record specific experiences over a period of time in order capture qualitative, contextualized user experiences, although quantitative data may also be logged automatically. These studies involve interventions in the form of semi-structured questions and optionally a prototype, but do not involve comparisons, although they do ask participants to reflect upon their experiences.

Speed Dating [36] is a design-oriented interventionist method created to assess design concepts and reflect on their merits by exposing users to multiple lo-fidelity sketches, storyboards or mock-ups of diverse design concepts. This method asks participants to quickly assess a relatively large number of early-stage designs, rather than the two to three that are common in *Structured Observation* with somewhat more advanced designs. Explicitly comparing the design concepts to each other is not required. Exposure to the design variants is not controlled according to the principles of experimental design, nor are tasks required.

Cultural probes [37] are a design-inspired HCI method intended to inspire ideas by creating artifacts and assigning evocative tasks that participants perform in specified settings. They do not seek to understand users in the traditional, social science sense, but instead seek to inspire ideas and stimulate the designer's imagination. The data is qualitative, with no comparisons, although participants may sometimes be asked to reflect upon their experiences.

Technology probes [38] ask users to try a new technology for a specified period of time in a field setting. This is an explicitly multi-method HCI approach that addresses three different types of research goals: obtaining qualitative and quantitative data about users' experiences; assessing technical constraints relative to the specific use setting; and inspiring ideas that advance the design concept. Technology probes involve intervention, both in the installation of the technology and in the form of post-hoc interviews. Participants are not asked to make comparisons but are asked to reflect upon their lived experiences.

Usability studies [23] are an HCI method inspired by Human Factors or Ergonomics research, with the goal of testing late-stage prototypes or products to determine if and how well users can access features to achieve their goals. The focus can be on either qualitative data about how the interactive technology was experienced or on quantitative performance data. Researchers intervene by asking users to perform pre-defined tasks with the prototypes/products. Although comparison is not required, participants may be asked to reflect on their preferences with respect to the system.

Quasi-experiments [18, 21], or more generally, field experiments, are drawn from psychology and other natural sciences, and seek to determine causal relationships. Although based on controlled experiments, they are designed to address potential threats to validity that arise when experimental conditions cannot be fully controlled in the field. Quasi-experiments focus on quantitative data generated by participants' performance on pre-defined tasks, although qualitative data may also be collected. Comparison is required although participant reflection is not.

Controlled experiments [39, 18] are drawn from psychology and other natural sciences, with the explicit goal of determining causal relationships across factors. Researchers intervene by defining initial

hypotheses that can be operationalized into participant exposure to conditions and well-defined tasks, with quantifiable measures of performance. Experiments focus on quantitative data and comparison is required although participant reflection is not.

The above research methods each address different research needs. *Comparative Structured Observation* provides a useful complement to these methods, and borrows characteristics from both qualitative and quantitative approaches. It combines the rigorous design of an interventionist study based on well-established experimental design principles with the careful analysis of participants' comparative reflections using well-known qualitative analysis methods. The goal is to organize the presentation of conditions so as to maximize the participant's ability to systematically compare and reflect upon relevant design characteristics and thus advance the design concept.

7 Characteristics of *Comparative Structured Observation*

Many HCI researchers face the problem of how to effectively characterize “mid-phase” research that occurs after initial work with users has led to a potentially interesting concept or design direction, but before it makes sense to ask precise, quantitatively testable research questions. Indeed, such questions may never make sense, especially when comparing and assessing complex, multi-faceted design concepts. In such cases, researchers are more interested in obtaining deep, grounded reflections from target users either about their experience with the new design compared to an existing system or among design variants within the new system design. As we have demonstrated, the published HCI literature already includes multiple examples of well-designed, qualitative comparative studies. However, we believe the lack of a clearly named qualitative comparison method with agreed-upon evaluation criteria has limited the number of such studies and encouraged authors to focus on quantitative results to the detriment of qualitative insights.

We characterize *Comparative Structured Observation* as follows:

Researchers explicitly define at least two comparable design variants that all participants experience as they perform the same or equivalent tasks. In experimental design terms, this means that the study must always include at least one within-participant factor. Design variants may include:

1. two or more variants of a novel design concept, e.g. designs for synchronous online lectures that add different types of simulated attendee proximity cues.
2. a new design and a baseline, where the baseline can be a different system, e.g. a novel vs. an established document editor, or the absence of the new system, e.g. a browser as the baseline vs. the browser with a novel plugin.
3. two or more variants of a novel design concept and a baseline, which simply combines (1) and (2) into a single study.

Design variants are usually expressed as medium-fidelity prototypes, although low-fi and hi-fi prototypes are possible. Designs must be sufficiently mature to capture how users will experience different aspects of the design concept through performing relevant tasks.

Researchers derive participant tasks based on formative research, either conducted directly by the researcher or derived from the research literature. Tasks are chosen to maximize ecological validity and generate qualitative data relevant to the design concept. When imposing tasks is difficult or irrelevant, the basis of comparison may be derived from an activity’s inherent structure, e.g., naturally occurring notifications that interrupt the user from a primary task.

Lab and field settings are both appropriate. Field studies offer greater ecological validity, but are not always possible, e.g., when a design implementation is not sufficiently robust for field use. Lab studies may allow participants to experience relevant events more quickly than under field conditions or confront them with rare but important events.

Participant exposure to tasks and design variants meets best practices of experimental or quasi-experimental design to mitigate known threats to validity.

Participants compare and reflect upon their experiences with each design variant. The study protocol includes capturing qualitative data where participants explicitly compare and reflect deeply on their experiences with the design variants.

Researchers compare and reflect on participants’ experiences based on direct observation or other sources of rich data. Ideally, observation is live and in person, but may be conducted remotely, e.g., via video link, or involve recorded video or other rich log data.

Researchers conduct post-hoc interviews, either semi-structured or open-ended, after participants have experienced the design variants. These interviews probe for additional information about partici-

pants' experiences or what researchers observed during the study, with the emphasis on comparing the different design variants.

Researchers emphasize gathering qualitative data to obtain rich, contextualized insights about design variants. This advances the goal of driving design innovation based on the grounded reflections reported by participants and the researcher's interpretations of their behavior and reflections.

Researchers may also capture quantitative data to help interpret qualitative data, but this is considered secondary. For example, the researcher may collect self-reported measures of satisfaction or measure time to complete a task and compare it to the user's perception of their own performance.

Researchers analyze qualitative data using recognized methods, such as reflexive thematic analysis [40] or grounded theory [41]. Thematic analyses will normally include both inductive and deductive analysis, since the structure of the participants' tasks or experiences will provide themes for some of the "top-down" analysis, and "bottom-up" insights are always welcome.

8 *Comparative Structured Observation* Criteria

Determining whether or not a particular study design should be considered a *Comparative Structured Observation* requires assessing it according to a specific set of criteria. Table 3 presents two checklists: the first is designed to help researchers determine whether or not a particular study design should be considered a *Comparative Structured Observation* study. The second suggests features for designing and conducting a “good” one.

Required CSO characteristics	Optional CSO characteristics	Incorrect CSO characteristics
Researchers explicitly define comparable design variants.	Lab and field settings are both appropriate.	Compares design to status quo only, outside the context of the study.
Researchers derive participant tasks based on formative research.	Participants use a talk aloud protocol.	Runs an open-ended field test with no comparisons.
Participant exposure to tasks and design variants meets experimental design best practices.	Quantitative data informs qualitative data.	Omits comparable tasks or experiences.
Participants experience at least two design variants within the study.	Researchers generate testable hypotheses.	Omits participant reflection on comparable experiences.
Participants compare and reflect upon their experiences with each design variant.		Omits researcher observation and reflection on participants' experiences.
Researchers observe participants' experiences directly or through other rich data sources.		Focuses only on performance metrics.
Researchers emphasize gathering qualitative data.		
Researchers conduct post-hoc interviews.		

Table 3: Left: *Comparative Structured Observation* checklist. Right: Suggested features of a “good” *Comparative Structured Observation* study.

Our goal is to improve the quality of future qualitative studies by encouraging both HCI researchers and reviewers to employ these criteria in a consistent way. Note that each criterion for what constitutes a good use of the method could be expanded into a full set of its own criteria. For example, best practice for reporting qualitative results includes a brief summary of the most “obvious” points, followed by a more detailed analysis of the most important or surprising findings; and best practice for creating an experimental design includes generating “equivalent” tasks and structuring their presentation to avoid unwanted nuisance effects, such as learning or fatigue. Unpacking each criterion is beyond the scope of the current article. We expect researchers who use *Comparative Structured Observation* to have sufficient prior training in both qualitative methods, e.g., [42, 40, 43, 44],

9 Applying *Comparative Structured Observation*

We have applied the *Comparative Structured Observation* approach to the design of several interactive AI applications. This deliverable describes how we used the approach to develop two systems that explore how to share agency between a user and an intelligent system. STORYLINES [45] concerns a safety critical system, specifically targeting control room operators at the French national power grid. *Design Prompt*[46]. targets creative professionals, specifically independent designers. Unfortunately, AML was not available in time for the development of these systems, so we used a more traditional machine learning methods to develop each system. However, each system demonstrates the benefits of using a participatory design approach, combined with *Comparative Structured Observation*, to design next-generation interactive intelligent systems.

9.1 StoryLines Project

Designing interactive technology to support safety-critical systems poses multiple challenges with respect to security, access to operators and the proprietary nature of the data. We were interested in how to take advantage of the increasing capabilities of artificial intelligence to assist control room operators of safety-critical systems and enhance the operator's skills over time. Here, the goal was to combine the capabilities of the human operator and the intelligent assistant such that together, they are more productive than either alone.

We conducted a two-year participatory design project with French power grid operators to both understand their specific needs and to collaborate on the design of a novel collaborative tool called STORYLINES. Our primary objective was to capture detailed, in-context data about operators' work practices as part of a larger project designed to provide bi-directional assistance between an intelligent agent and human operator. We targeted handovers between shifts to take advantage of the operators' existing practice of articulating the current status of the grid and expected future events. We use information that would otherwise be lost to gather valuable information about the operator's decision rationale and decision-making patterns. We combined a bottom-up participatory design approach with a top-down generative theory approach [11] to design STORYLINES, an interactive timeline that helps operators collect information from diverse tools, record reminders and share relevant information with the next shift's operator.

9.2 StoryLines Study

We wanted to combine our theoretical work on generative theory with a participatory design approach. As we explained earlier, the traditional "user-centered" approach to designing interactive systems can be oversimplified as first conducting a user study to determine users' needs, then exploring a new design and finally testing it with target users. By contrast, participatory design actively involves users throughout the design process. While this has proven to be highly successful in a few safety-critical settings, such as that described by Mackay [47, 48] with air traffic controllers, it is relatively rare due to security restrictions, limited physical access to users and their work environments, and the confidential or proprietary nature of the technologies and associated data.

We also cannot risk losing participants' interest by forcing them to adopt our theoretical stance. Instead, we need to let them express their experiences and needs in their own terms while contributing concrete examples of design ideas inspired by our theories which they can interpret and evaluate as they like.

We followed the participatory design approach outlined in [49, 6, 7], beginning with story interviews that capture detailed examples of recent, memorable incidents, especially breakdowns, workarounds and user innovations. We transformed these into scenarios in which personas who represent actual controllers encounter a series of events that highlight the issues that must be resolved. We brainstormed ideas, first verbally then as short video clips that illustrate the interaction [50] and then collaborated on creating various types of prototypes [51], ranging from paper to Figma to working software. We ran different types

of qualitative evaluations to advance the design, including generative walkthroughs [52]. and iterated with additional interviews, brainstorming, design sessions and evaluations. Participants found this approach highly engaging, since all the activities were designed to emphasize their perspective, with concrete activities that let them express their needs and ideas, as well as react to our ideas, in an enjoyable, non-threatening format.

9.2.1 Participants:

We recruited a total of four active participants who contributed to multiple activities over several months and three additional operators who contributed to specific events. One very senior operator participated throughout the entire design process. The junior operator who participated in the first series of activities left RTE after the first year. He was replaced by a junior and a senior operator who were actively involved in later activities. Three other operators participated in individual workshops and others contributed information or joined discussions at various points in the process.

9.2.2 Control room organization:

Four researchers visited the RTE control center which began with a “virtual” guided tour from the senior operator. The main control room is designed to support four operators: three who supervise a particular zone (east, west and south) and a senior manager who oversees the others. During busy times, a fifth operator may also be included. Operators are trained on all three zones and spend approximately seven weeks in one zone before shifting to another. The manager supervises all three zones.

The three operator desks were positioned to face a wall-sized display on the back wall with a huge map that displayed the entire energy grid. The manager’s desk is positioned behind them, with a view of all three desks. Each operator has a large U-shaped desk with an interior approximately three meters wide. We counted about a dozen “stations” for different tools and information sources, including multiple screens, a printer, two telephone stations and areas for other devices. One set of three screens is operated with a single mouse and keyboard, whereas other tools have dedicated keyboards or control panels. Some tools provide overviews and details about the current status of the energy grid, some sound and monitor alarms while others help operators research re-routing solutions to current failures, expected maintenance or other projected problems. The size of the desk area requires them to use rolling chairs to transition from one station to another.

Software tools. Energy grid operators are trained on over 50 different software tools that gather information, sound alarms and suggest solutions. These tools have been introduced over several decades with correspondingly different, often incompatible user interfaces. Operators must collect information scattered across their work area and they struggle to form a coherent overview of the power grid. Most information is presented as raw data that rarely matches the operators’ mental models. Worse, some information is cumbersome or difficult to retrieve in real-time and junior operators reported that instructions and guidelines were often difficult to understand.

Paper-based tools. Operators also use multiple paper-based tools. An A0-sized paper map of the energy grid, reprinted monthly, is used to note current breakdowns which they must remember to erase when the problem is solved. Although this paper map provides a concise overview of the grid, it is usually out of date since it is not linked to any online software tools. Other paper-based supports include a large ring binder, exchanged monthly, with hand-annotated printouts of specific “incidents”, an A3-sized log sheet for recording completed maneuvers and “floating sheets” for informally recording notes, reminders and information for the shift handover.

We were particularly interested in the fact that all operators took paper notes that they used to keep track of ongoing activities and manually transfer information from one system to another (which they view as

a waste of time). They also create reminders for future actions and list key items to communicate to the next shift's operator during the handover. We were also struck by the level of oral communication between operators, not only during shift handovers but also across zones and with the manager. Operators were often interrupted by phone calls from clients and other stakeholders. From the perspective of the CAB project, this is extremely valuable information that is lost and unavailable to the system.

9.2.3 Participatory design workshops:

After the preliminary study, we conducted a series of participatory design workshops over a period of one year. Figure 1 shows this sequence of activities with operators, both in person and online, as well as the evolution of the STORYLINES prototype from paper mock-ups to working software.

Workshop 2. The success of workshop 1 led to an invitation to visit the RTE control center in Nantes. Workshop 2 was an all-day event with seven participants (three controllers and four researchers). Six were physically present and one controller joined via video from the control room. The controllers gave us a virtual tour of the control room (described above) after which we brainstormed ideas for how to address the breakdowns specified in the scenarios from workshop 1. Next, we paired controllers with researchers to draw a short “interaction snippet” or storyboard to illustrate each idea. Finally, each pair created paper mock-ups and shot short video clips where the controllers demonstrated how they would interact with each idea. The operators especially liked the video brainstorming activity because it gave them a concrete way to communicate both the problem and the solution. The workshop resulted in 10 interaction snippets and five video prototypes, all of which highlighted the need for centralizing information.

Workshop 3. The third workshop was also conducted in person, in Nantes and focused on situating the breakdowns identified earlier in a realistic context. We began with a visit to the observation deck above the actual control room, which allowed us to observe the operators in action. Next, we ran a design session that involved creating realistic personas based on the operators involved in the project and a “current scenario” to illustrate the breakdowns in a realistic context. The operators drew from their experiences to adjust the scenarios, especially with respect to the ways in which small errors, particularly in communication, can cause problems or complicate the situation.

The workshop resulted in scenarios related to two personas: Charles and Julie. Charles is a senior operator with 12 years of experience and Julie is a junior operator who just finished training. The scenario includes three main situations: a busy morning for Charles who just returned from vacation; a rushed and poorly prepared handover to Julie and an afternoon filled with complex problems for Julie.

Workshop 4. The fourth workshop was conducted in our labs with five researchers and two operators. We developed a “future scenario” based on the scenario created in workshop 3 that explored potential solutions. We supplied prototyping materials and simulated their environment with paperboard “screens”. We then created paper mock-ups of the different tools and created video prototypes of various solutions proposed by the operators and researchers. The results included four specific solutions designed to improve communication with the planning team, visualize cognitive load and visualize failures as they evolve over time.

Workshop 5. The fifth workshop was conducted at the RTE control center with three researchers and one operator physically present as well as one remote senior operator. The goal of the workshop was to use a generative walk-through [52] to evaluate the scenarios and ideas from workshop 4. We played the video prototypes and discussed how to improve the proposed solutions, including incorporating new information as a situation evolves, automatic update of other information and tasks, tracking and comparing solutions and taking into account the “last time to decide” (LLTD). This workshop highlighted

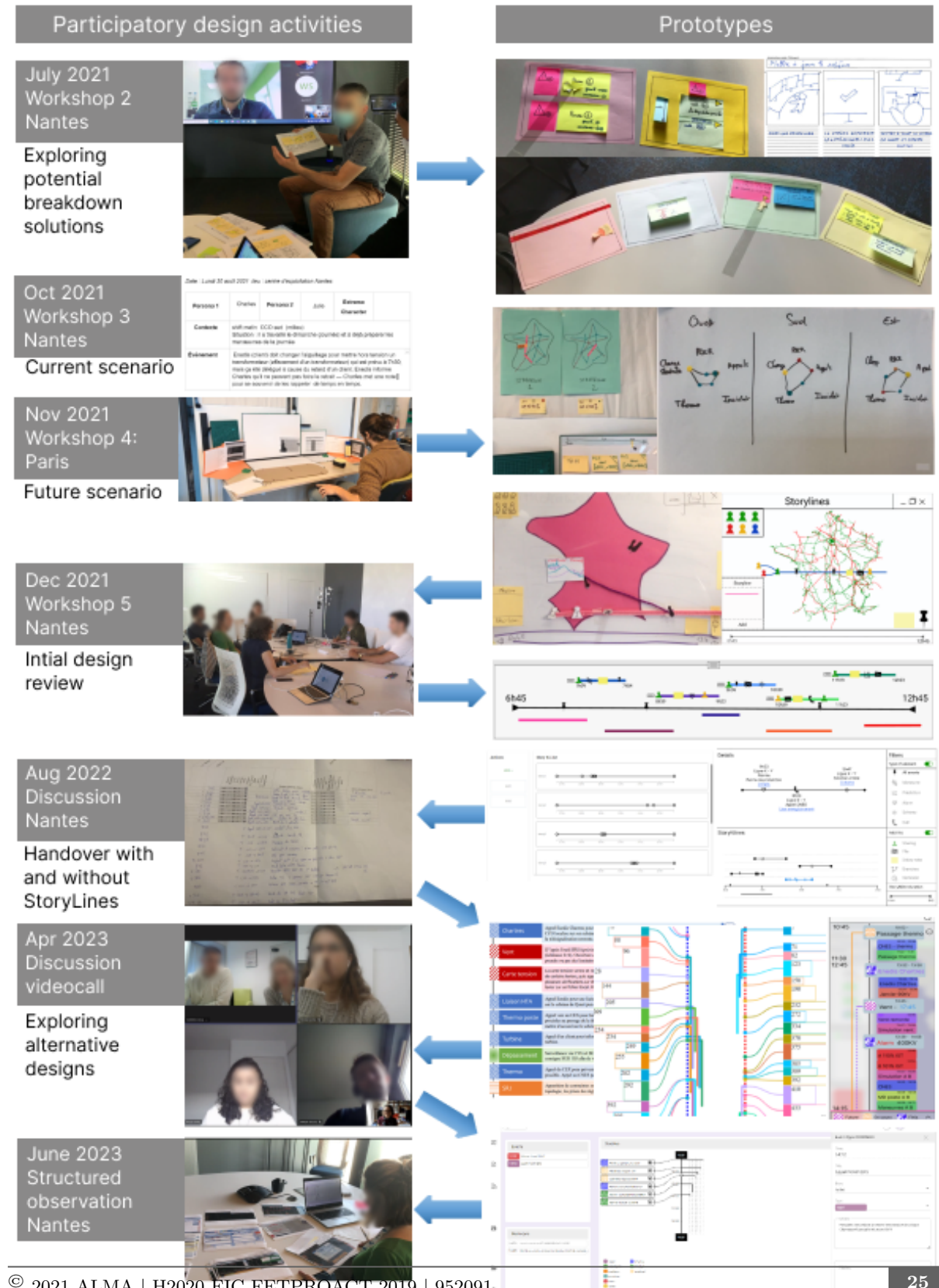


Figure 1: Participatory design activities alternated between design sessions with operators (left column) and prototypes created by operators and researchers (right column) over a period of approximately two years.

how much critical information is lost, either because it was transmitted orally or it was never captured in the first place.

Workshop 6. The sixth workshop was also conducted at the RTE control center with four researchers and five operators. Our goal was to present the operators with a working prototype of STORYLINES and obtain their feedback, suggestions and ideas. We worked with pairs of operators who followed two scenarios, one extremely busy with multiple overlapping events and the other relatively calm. They performed tasks, interacted with STORYLINES, took notes and prepared for the handover to the other operator, after which they continued with the second scenario. Half of the operators first performed the “busy” scenario, followed by the “calm” one. The other half performed the “calm” scenario first. We conducted two sessions with two operators each and then ran a debriefing session with the operators who participated plus another senior controller.

The participatory design approach enabled us to actively involve operators in the design process, who have been highly engaged and enthusiastic about the project (in contrast to a parallel project that was conducted in a more traditional, top-down manner). We will be continuing this approach to further refine the prototype, with the longer-term goal of further developing the system into an open-source project both for these users and others facing similar challenges.

9.3 DesignPrompt Project

A second project targets visually oriented designers, who often struggle to create effective generative AI (GenAI) prompts. A preliminary study identified specific issues in composing and fine-tuning prompts, as well as needs in accurately translating intentions into rich input. We developed DESIGNPROMPT, a moodboard tool that lets designers combine multiple modalities — images, color, text — into a single GenAI prompt and tweak the results. We ran a comparative structured observation study with 12 professional designers to better understand their intent expression, expectation alignment and transparency perception using DESIGNPROMPT and text input GenAI. We found that multimodal prompt input encouraged designers to explore and express themselves more effectively. Designer’s interaction preferences change according to their overall sense of control over the GenAI and whether they are seeking inspiration or a specific image. Designers developed innovative uses of DESIGNPROMPT, including developing elaborate multimodal prompts and creating a multimodal prompt pattern to maximize novelty while ensuring consistency.

We conducted a comparative structured observation study that systematically controlled the presentation of the user’s *prompt strategy*. Participants performed multiple tasks with each design variant and then respond to questions that require them to reflect deeply upon and compare their strengths and weaknesses. An experimenter also directly observes how participants interact with each design variant. The goal is to understand the trade-offs across design variants from the user’s perspective to evaluate and advance the design.

To provide a fair comparison, we created two variations of DESIGNPROMPT, a moodboard creation system that differed only in the GenAI prompt interaction: The *Text condition* permits only text-based prompts and basic inpainting, similar to DALL-E’s AI image generation interface, and serves as the baseline. The *Multimodal condition* permits multimodal prompts, including the above as well as images and color, and serves as the key design variant. At this writing, current moodboard tools do not integrate GenAI prompts and inpainting with an integrated search engine, and thus our baseline exceeds current capabilities. However, although this raises the bar for the baseline comparison, we also anticipate that these capabilities will arrive soon and so consider this a fairer comparison.

9.4 Participants

We recruited 12 participants (9 women, 3 men; aged 22-54) who have at least two years of design training or practice; and 2-20 years of moodboard design experience. Their experience with GenAI varied from novice to expert. We followed our institution's IRB policy both in our explanations of the study to participants and study design. All participants signed informed consent forms and agreed to voice and screen recordings. We anonymized the results and followed European privacy law (GDPR).

9.5 Results

The short questionnaires at the end of each design task showed that both the *The Text condition* and *The Multimodal condition* prompt strategies improved system expressivity, users' understanding of prompt and images' connection, and output understanding, and are less good at producing images that "align with expectations". This indicates that both the *The Text condition* and *The Multimodal condition* prompt strategies are suitable for moodboard design.

The results of the comparative questionnaire (Fig.2) showed that participants preferred the *The Multimodal condition* prompt strategy for exploration, expressivity, interaction worth the effort and user understanding on how prompts impact AI. However, they preferred the *The Text condition* prompt strategy for ease of steering and generating images that makes sense. Participants preferred the *The Multimodal condition* strategy (41%) to *The Text condition* (25%) to support ideation, but were mixed with respect to "feeling in control", with roughly equal numbers of participants preferring each.

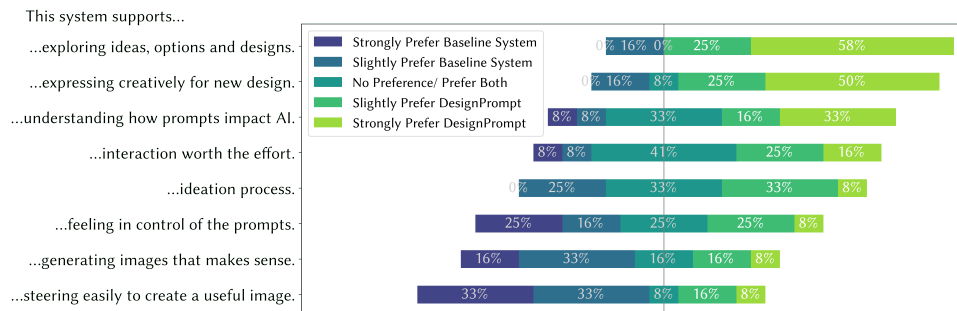


Figure 2: Comparative Questionnaire Results between the *Text condition* (blue) and *The Multimodal condition* (green).

Multimodal input enabled designers to explore and express their ideas, without feeling as though the system was in control. Most designers mentioned they enjoyed the multimodal input feature (8/12), especially in terms of expressing themselves more creatively (9/12) exploring different ideas (10/12) and providing them more options and opportunities (4/12). Participants enjoyed the creative options and opportunities *multimodal condition* provided (P2, P4, P7, P10). P1 also mentioned in the *multimodal condition* "was much better in terms of workflow" and "(make moodboarding) more enjoyable to do". The paper on *DesignPrompt* [46] provides a detailed summary of additional qualitative findings.

10 Discussion

As demonstrated in section 9, participatory design offers an effective approach for exploring innovative forms of interaction with intelligent systems. This approach lets us focus on the details of the *interaction* between the user and the AI: rather than treating the user a component in an intelligent system, we need to switch the focus to the overall human system, with intelligent agents acting as an important, but not primary component.

Note that this flies in the face of the traditional “human-in-the-loop” approach, which treats the user as a source of input to the algorithm, like any other. In our experience, AI researchers treat users as simply sources of data for their algorithms. They rarely consider what the experience of providing data to the computer is like: Is it a boring or interesting activity? Is it perceived as useful or a waste of time? Does it matter if the user makes a mistake, and if so, how can they correct it? Is it clear how the data will be used and if that user will obtain any benefit, either directly or later?

These questions all arise when we consider the user not as another cog in the machine, but rather as a sentient being with different skills, values and assumptions from the computer.

Another common assumption from the AI researcher is that the user is a source of unwanted error. The most common reaction to this is to further limit the scope of the user’s activity, which make limit potential error, but also makes the act of inputting data significantly more boring and repetitive, which may itself increase the likelihood of error.

Users are rarely treated as equal partners, with equal or greater autonomy than the AI system. Of course, providing users with greater agency and flexibility is difficult. Users are unpredictable and systems are easier to build if the input source is predictable.

However, if we switch the perspective from treating the user as the “human-in-the-loop”, to a more open-ended approach where the AI algorithm is simply the “computer-in-the-loop” that is controlled by the user, we can start to design intelligent systems that make sense in human terms, for human users. If we value the fact that users do in fact think and can respond to novel situations with novel solutions, the overall combination of the human user and the AI system will benefit.

11 Conclusion

This report corresponds to *Deliverable D4.6 - Interaction Paradigm Methodology - Final* of the ALMA project. The goal of this deliverable is to extend current participatory design and generative design methodologies to incorporate the role of an intelligent system in the interaction loop, where agency shared between the human user and the AML system.

We first describe the challenges associated with the wide-spread introduction of machine learning and other intelligent systems into interactive systems. *Deliverable 4.5* goes into depth about the problems associated with incorporating AI in ways that deskill or replace the user. We argue that a key role for HCI is to demonstrate how to change the design process so that the users' needs and values are respected, and that the resulting systems actively benefit users rather than treat them as simply sources of data, or sources of error.

The next section describes how HCI has changed as a field to accommodate different waves of technology and the associated increased penetration of computers into all aspects of our work and personal lives. It describes the key characteristics of participatory design, with an emphasis on the values it places and the different strategies it uses for including users in the design process. We then describe how our new theoretical approach on *generative theories of interaction* allows us to benefit from theories developed in other disciplines and transform them into practical design principles that can be applied to non-AI and AI-based systems alike. Next, we present *Comparative Structured Observation*, a mixed-method qualitative approach that provides designers with a rigorous method for gathering grounded, comparative reflections on design alternatives from users. We briefly discuss the benefits of this approach and conclude with directions for future research.

References

- [1] Janet Rafner et al. “Deskillling, Upskilling, and Reskilling: a Case for Hybrid Intelligence”. In: *Morals & Machines* 1 (Jan. 2021), pp. 24–39. DOI: 10.5771/2747-5174-2021-2-24.
- [2] Naomi Klein. “AI machines aren’t ‘hallucinating’. But their makers are”. In: *The Gardian* (May 2023).
- [3] Susanne Bødker. “Third-Wave HCI, 10 Years Later—Participation and Sharing”. In: *Interactions* 22.5 (Aug. 2015), pp. 24–31. ISSN: 1072-5520. DOI: 10.1145/2804405. URL: <https://doi.org/10.1145/2804405>.
- [4] Clay Spinuzzi. “The Methodology of Partioipatory Design”. In: *Society for Technical Communication: Applied Research* 52.2 (May 2005), pp. 163–174.
- [5] Agile Alliance. *What is Agile Software Development?* Tech. rep. Agile Alliance, 2013.
- [6] Wendy E. Mackay. “Designing with Sticky Notes”. In: *Sticky Creativity: Post-It Note Cognition, Interaction and Digitalization*. Ed. by Bo Christensen, Kim Halskov, and Clemens Klokmoose. Cambridge, MA, USA: Academic Press, Nov. 2019, pp. 231–256. URL: <https://hal.science/hal-02426827>.
- [7] Wendy E. Mackay. *DOIT: The Design of Interactive Things. CHI 2023 Preview*. Paris, France: Inria, Apr. 2023.
- [8] Michel Beaudouin-Lafon and Wendy E. (Mackay. “Prototyping Tools and Techniques”. In: *Handbook of Human-Computer Interaction, revised edition*. Ed. by A. Jacko and A. Sears. New York, New York: Lawrence Erlbaum Associates, 2007, pp. 1006–1031.
- [9] Wendy E. Mackay and Anne-Laure Fayard. “HCI, Natural Science and Design: A Framework for Triangulation Across Disciplines”. In: *Proceedings of the 2nd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*. DIS ’97. Amsterdam, The Netherlands: ACM, 1997, pp. 223–234. ISBN: 0-89791-863-0. DOI: 10.1145/263552.263612. URL: <http://doi.acm.org/10.1145/263552.263612>.
- [10] Wendy E. Mackay and Joanna McGrenere. “Comparative Structured Observation”. In: *ACM Transactions on CHI* (2025), p. 26.
- [11] Michel Beaudouin-Lafon, Susanne Bødker, and Wendy E. Mackay. “Generative Theories of Interaction”. In: *ACM Trans. Comput.-Hum. Interact.* 28.6 (Nov. 2021). ISSN: 1073-0516. DOI: 10.1145/3468505. URL: <https://doi.org/10.1145/3468505>.
- [12] Wendy Mackay. “Responding to cognitive overload: Co-adaptation between users and technology”. eng. In: *Intellectica* 30.1 (2000), pp. 177–193. ISSN: 0769-4113. DOI: 10.3406/intel.2000.1597.
- [13] Michel Beaudouin-Lafon. “Instrumental Interaction: An Interaction Model for Designing post-WIMP User Interfaces”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’00. The Hague, The Netherlands: ACM, 2000, pp. 446–453. ISBN: 1-58113-216-6. DOI: 10.1145/332040.332473. URL: <http://doi.acm.org/10.1145/332040.332473>.
- [14] Michel Beaudouin-Lafon and Wendy E. Mackay. “Reification, Polymorphism and Reuse: Three Principles for Designing Visual Interfaces”. In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI ’00. Palermo, Italy: ACM, 2000, pp. 102–109. ISBN: 1581132522. DOI: 10.1145/345513.345267. URL: <https://doi.org/10.1145/345513.345267>.
- [15] Saul Greenberg and Bill Buxton. “Usability Evaluation Considered Harmful (Some of the Time)”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’08. Florence, Italy: Association for Computing Machinery, 2008, pp. 111–120. ISBN: 9781605580111. DOI: 10.1145/1357054.1357074. URL: <https://doi.org/10.1145/1357054.1357074>.
- [16] Dan R. Olsen. “Evaluating User Interface Systems Research”. In: *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*. UIST ’07. Newport, Rhode Island, USA: Association for Computing Machinery, 2007, pp. 251–258. ISBN: 9781595936790. DOI: 10.1145/1294211.1294256. URL: <https://doi.org/10.1145/1294211.1294256>.
- [17] Henry Lieberman. *The Tyranny of Evaluation*. New York, NY, USA: ACM, 2003. URL: <https://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html>.

- [18] Thomas D Cook and D T Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. English. New York: Houghton Mifflin, 1979.
- [19] C. Lewis. *Using the "thinking Aloud" Method in Cognitive Interface Design*. Research report RC-9265. IBM Thomas J. Watson Research Division. URL: <https://books.google.fr/books?id=F5AKHQAACAAJ>.
- [20] Hannu Kuusela and Pallab Paul. "A Comparison of Concurrent and Retrospective Verbal Protocol Analysis". In: *The American Journal of Psychology* 113.3 (2000), pp. 387–404. ISSN: 00029556. URL: <http://www.jstor.org/stable/1423365> (visited on Sept. 1, 2024).
- [21] W. R. Shadish, T. D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2nd ed. New York: Houghton Mifflin, 2001.
- [22] R. Johnson and Anthony Onwuegbuzie. "Mixed Methods Research: A Research Paradigm Whose Time Has Come". In: *Educational researcher* 33 (Oct. 2004), p. 14. DOI: 10.3102/0013189X033007014.
- [23] Jakob Nielsen. "Estimating the number of subjects needed for a thinking aloud test". In: *International Journal of Human-Computer Studies* 41.3 (1994), pp. 385–397. ISSN: 1071-5819. DOI: <https://doi.org/10.1006/ijhc.1994.1065>.
- [24] Wayne D. Gray and Marilyn C. Salzman. "Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods". In: *Human-Computer Interaction* 13.3 (1998), pp. 203–261. DOI: 10.1207/s15327051hci1303_2.
- [25] P.J. Runkel and J.E. McGrath. *Research on Human Behavior: A Systematic Guide to Method*. New York, NY, USA: Holt, Rinehart and Winston, 1972. ISBN: 9780030786952. URL: <https://books.google.dk/books?id=3FdqAAAAAAAJ>.
- [26] Joseph E. McGrath. "Methodology Matters: Doing Research in the Behavioral and Social Sciences". In: *Readings in Human-Computer Interaction*. Ed. by Ronald M. BAECKER et al. Interactive Technologies. New York: Morgan Kaufmann, 1995, pp. 152–169. ISBN: 978-0-08-051574-8. DOI: <https://doi.org/10.1016/B978-0-08-051574-8.50019-4>. URL: <https://www.sciencedirect.com/science/article/pii/B9780080515748500194>.
- [27] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [28] John W. Tukey. "We Need Both Exploratory and Confirmatory". In: *The American Statistician* 34.1 (1980), pp. 23–25. ISSN: 00031305. URL: <http://www.jstor.org/stable/2682991> (visited on July 4, 2022).
- [29] J. Greenbaum and M. (Eds.) Kyng. *Design at Work: Cooperative Design of Computer Systems (1st ed.)*. Boca Raton, Florida, USA: CRC Press, 1991, p. 306.
- [30] Karl Lewin. "Frontiers in Group Dynamics". In: *Human Relations* 1 (1947), pp. 5–41. DOI: 10.1177/001872674700100103.
- [31] Susanne Bødker. *Through the Interface: A Human Activity Approach to User Interface Design*. USA: L. Erlbaum Associates Inc., 1991. ISBN: 0805805702.
- [32] M.V. Angrosino. *Naturalistic Observation (1st ed.)*. New York: Routledge, 2007. DOI: 10.4324/9781315423616.
- [33] Hugh Beyer and Karen Holtzblatt. *Contextual Design: Defining Customer-Centered Systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. ISBN: 9780080503042.
- [34] Csikszentmihalyi M and Larson R. "Validity and reliability of the Experience-Sampling Method". In: *J Nerv Ment Dis*. 175.9 (Sept. 1987), pp. 526–36. DOI: 10.1097/00005053-198709000-00004.
- [35] John Rieman. "The Diary Study: A Workplace-Oriented Research Tool to Guide Laboratory Efforts". In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*. CHI '93. Amsterdam, The Netherlands: Association for Computing Machinery, 1993, pp. 321–326. ISBN: 0897915755. DOI: 10.1145/169059.169255. URL: <https://doi.org/10.1145/169059.169255>.
- [36] Scott Davidoff et al. "Rapidly Exploring Application Design Through Speed Dating". In: *UbiComp 2007: Ubiquitous Computing*. Ed. by John Krumm et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 429–446.

- [37] Bill Gaver, Tony Dunne, and Elena Pacenti. “Design: Cultural Probes”. In: *Interactions* 6.1 (1999), pp. 21–29. ISSN: 1072-5520. DOI: 10.1145/291224.291235. URL: <https://doi.org/10.1145/291224.291235>.
- [38] Hilary Hutchinson et al. “Technology Probes: Inspiring Design for and with Families”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. Ft. Lauderdale, Florida, USA: ACM, 2003, pp. 17–24. ISBN: 1-58113-630-7. DOI: 10.1145/642611.642616. URL: <http://doi.acm.org/10.1145/642611.642616>.
- [39] Scott E Maxwell, Harold D Delaney, and Ken Kelley. *Designing experiments and analyzing data: A model comparison perspective*. Abingdon-on-Thames, England, UK: Routledge, Taylor & Francis Group, 2017, p. 1080.
- [40] Virginia Braun and Victoria Clarke. “Reflecting on reflexive thematic analysis”. In: *Qualitative Research in Sport, Exercise and Health* 11.4 (2019), pp. 589–597. DOI: 10.1080/2159676X.2019.1628806.
- [41] Anselm L. Strauss and Juliet M. Corbin. *Basics of qualitative research: techniques and procedures for developing grounded theory*. Thousand Oaks, California: Sage Publications, 1998, p. 312.
- [42] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa.
- [43] Juliet M Corbin and Anselm Strauss. “Grounded theory research: Procedures, canons, and evaluative criteria”. In: *Qualitative sociology* 13.1 (1990), pp. 3–21.
- [44] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, California: Sage Publications, 2006.
- [45] *Creating StoryLines: Participatory Design with Power Grid Operators*. Springer, 2023, pp. 212–230.
- [46] Xiaohan Peng, Janin Koch, and Wendy E Mackay. “DesignPrompt: Using Multimodal Interaction for Design Exploration with Generative AI”. In: *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. ACM. Copenhagen, Denmark, July 2024, pp. 1–15. DOI: 10.1145/3643834.3661588. URL: <https://hal.science/hal-04652020>.
- [47] Wendy Mackay. “Responding to cognitive overload : Co-adaptation between users and technology”. eng. In: *Intellectica* 30.1 (2000), pp. 177–193. ISSN: 0769-4113. DOI: 10.3406/intel.2000.1597.
- [48] Wendy E. Mackay et al. “Reinventing the Familiar: Exploring an Augmented Reality Design Space for Air Traffic Control”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '98. Los Angeles, California, USA: ACM Press/Addison-Wesley Publishing Co., 1998, pp. 558–565. ISBN: 0201309874. DOI: 10.1145/274644.274719.
- [49] Wendy Mackay. “Using video to support interaction design”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '02. Minneapolis, Minnesota, USA: ACM, 2002. ISBN: 1-58113-516-5.
- [50] Wendy E Mackay, Anne V Ratzer, and Paul Janecek. “Video artifacts for design: bridging the Gap between abstraction and detail”. In: *DIS '00*. ACM, Aug. 2000, pp. 72–82. ISBN: 978-1-58113-219-9.
- [51] Wendy E. Mackay and Michel Beaudouin-Lafon. “Participatory Design and Prototyping”. In: *Handbook of Human Computer Interaction*. Springer International Publishing, Feb. 2023. DOI: 10.1007/978-3-319-27648-9_31-1. URL: <https://hal.science/hal-04108636>.
- [52] Danielle Lottridge and Wendy E Mackay. “Generative walkthroughs: to support creative redesign”. In: *Proceedings of the seventh ACM conference on Creativity and cognition*. 2009, pp. 175–184.