Reciprocal Explanations: An Explanation Technique for Human-AI Partnership in Design Ideation

Lena Hegemann

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo, April 27, 2020

Thesis Supervisor at Aalto University

Prof. Antti Oulasvirta

Thesis Examiner at KTH Royal Institute of Technology

Prof. Haibo Li

Thesis Advisors

M.Sc. Alexander Finn

Dr. Marianela Ciolfi Felice

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Lena Hegemann

Title: Reciprocal Explanations: An Explanation Technique for Human-AI Partnership in Design Ideation

Date: April 27, 2020 Language: English Number of pages: 9+61

Department of Computer Science

Professorship: User Interfaces

Supervisor: Prof. Antti Oulasvirta

Advisors: M.Sc. Alexander Finn, Dr. Marianela Ciolfi Felice

Advancements in creative artificial intelligence (AI) are leading to systems that can actively work together with designers in tasks such as ideation, i.e. the creation, development, and communication of ideas. In human group work, making suggestions and explaining the reasoning behind them as well as comprehending other group member's explanations aids reflection, trust, alignment of goals and inspiration through diverse perspectives. Despite their ability to inspire through independent suggestions, state-of-the-art creative AI systems do not leverage these advantages of group work due to missing or one-sided explanations. For other use cases, AI systems that explain their reasoning are already gathering wide research interest. However, there is a knowledge gap on the effects of explanations on creativity. Furthermore, it is unknown whether a user can benefit from also explaining their contributions to an AI system. This thesis investigates whether reciprocal explanations, a novel technique which combines explanations from and to an AI system, improve the designers' and AI's joint exploration of ideas. I integrated reciprocal explanations into an AI aided tool for moodboard design, a common method for ideation. In our implementation, the AI system uses text to explain which features of its suggestions match or complement the current moodboard. Occasionally, it asks for user explanations providing several options for answers that it reacts to by aligning its strategy. A study was conducted with 16 professional designers who used the tool to create moodboards followed by presentations and semi-structured interviews. The study emphasized a need for explanations that make the principles of the system transparent and showed that alignment of goals motivated participants to provide explanations to the system. Also, enabling users to explain their contributions to the AI system facilitated reflection on their own reasons.

Keywords: Interactive Machine-Learning, Collaborative AI, Design Ideation, Creativity Support Tools, Explainable AI

AALTO-UNIVERSITETET
HÖGSKOLAN FÖR TEKNIKVETENSKAPER

SAMMANDRAG AV
DIPLOMARBETET

Författare: Lena Hegemann

Titel: Ömsesidiga Förklaringar: En förklaringsteknik för Human-AI-samarbete inom konceptutveckling

Datum: April 27, 2020 Språk: Engelska Sidantal: 9+61

Institutionen för datateknik

Professur: Användargränssnitt

Övervakare: Prof. Antti Oulasvirta

Handledare: M.Sc. Alexander Finn, Dr. Marianela Coilfi Felice

Framsteg inom kreativ artificiell intelligens (AI) har lett till system som aktivt kan samarbeta med designers under idéutformningsprocessen, dvs vid skapande, utveckling och kommunikation av idéer. I grupparbete är det viktigt att kunna göra förslag och förklara resonemanget bakom dem, samt förstå de andra gruppmedlemmarnas resonemang. Detta ökar reflektionsförmågan och förtroende hos medlemmarna, samt underlättar sammanjämkning av mål och ger inspiration genom att höra olika perspektiv. Trots att system, baserade på kreativ artificiell intelligens, har förmågan att inspirera genom sina oberoende förslag, utnyttjar de allra senaste kreativa AI-systemen inte dessa fördelar för att facilitera grupparbete. Detta är på grund av AI-systemens bristfälliga förmåga att resonera över sina förslag. Resonemangen är ofta ensidiga, eller saknas totalt. AI-system som kan förklara sina resonemang är redan ett stort forskningsintresse inom många användningsområden. Dock finns det brist på kunskap om AI-systemens påverkan på den kreativa processen. Dessutom är det okänt om en användare verkligen kan dra nytta av möjligheten att kunna förklara sina designbeslut till ett AI-system. Denna avhandling undersöker om ömsesidiga förklaringar, en ny teknik som kombinerar förklaringar från och till ett AI system, kan förbättra designerns och AI:s samarbete under utforskningen av idéer. Jag integrerade ömsesidiga förklaringar i ett AI-hjälpmedel som underlättar skapandet av stämningsplank (eng. moodboard), som är en vanlig metod för konceptutveckling. I vår implementering använder AI-systemet textbeskrivningar för att förklara vilka delar av dess förslag som matchar eller kompletterar det nuvarande stämningsplanket. Ibland ber den användaren ge förklaringar, så den kan anpassa sin förslagsstrategi efter användarens önskemål. Vi genomförde en studie med 16 professionella designers som använde verktyget för att skapa stämningsplank. Feedback samlades genom presentationer och semistrukturerade intervjuer. Studien betonade behovet av förklaringar och resonemang som gör principerna bakom AI-systemet transparenta för användaren. Höjd sammanjämkning mellan användarens och systemets mål motiverade deltagarna att ge förklaringar till systemet. Genom att göra det möjligt för användare att förklara sina designbeslut för AI-systemet, förbättrades också användarens reflektionsförmåga över sina val.

Nyckelord: Interaktiv maskininlärning, samarbetande AI-agenter, idéutveckling, kreativitet stödjande verktyg, medskapande datortillämpningar, Förklarbar AI

Preface

First, I want to thank Professor Antti Oulasvirta, who supervised my thesis at Aalto University, and Professor Haibo Li, who was my examiner at KTH Royal Institute of Technology, for their guidance and feedback during the creation of this thesis. Furthermore, thank you to my advisor Dr. Marianela Coilfi Felice from KTH for her actionable feedback and organizing regular group discussions with other students working on their theses. Thank you also to my advisor Alexander Finn for his guidance during my internship at Silo AI. Furthermore, thank you to Janin Koch and Antti Oulasvirta for giving me the exciting opportunity to work with the May AI tool and the supervision and feedback during the design and implementation phase. Special thanks also go to Robert Talling and Daniel Wärnå who helped me with the translation of my abstract page to Swedish and to Michael Hedderich, Niels Dikken, Márton Elodi, and Jennifer Wang Kurtto for tips and pointers and their feedback to my manuscript. Last but not least I would like to thank all friends and family members who gave me moral support during my master's studies.

Espoo, April 27, 2020

Lena Hegemann

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
CST	Creativity Support Tool
HCI	Human-Computer Interaction
MB	Mood Board
RE	Reverse Explanation
SE	System Explanation

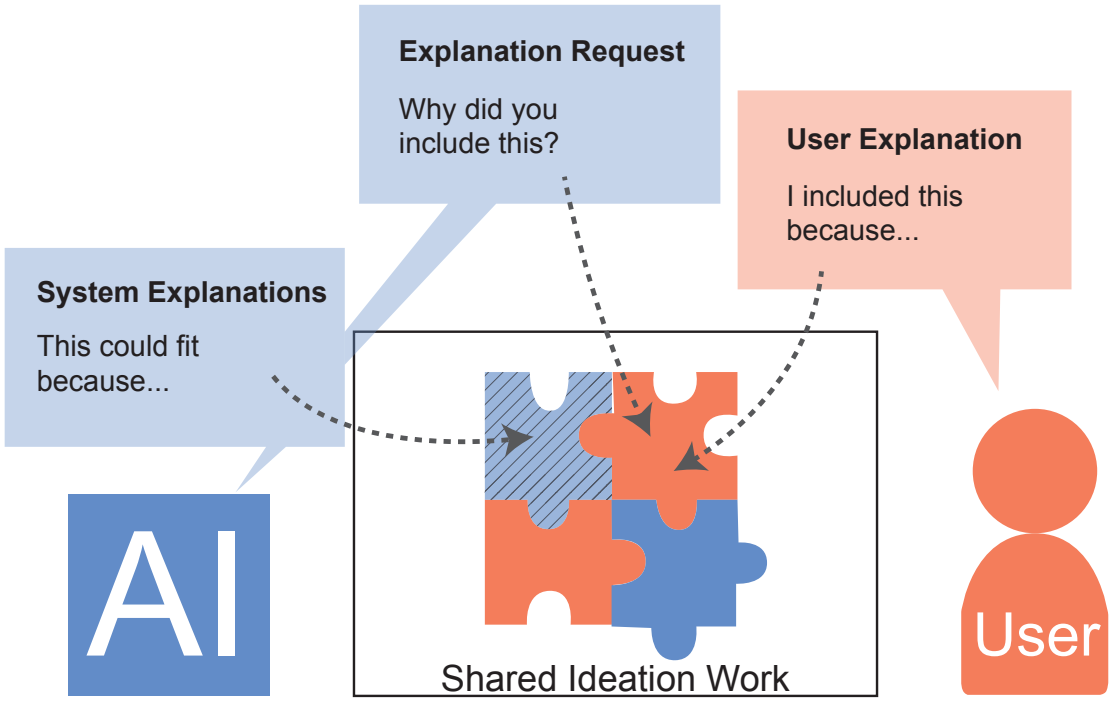

Figure 1: Reciprocal explanations are an explanation technique for scenarios where a human user and an AI system, both contribute ideas to a shared piece of ideation work. User contributions are depicted in red, AI contributions in blue. Crosslines indicate a suggestion. Reciprocal explanations allow for either partner to provide the reasons for their contributions (system explanations and user explanations). Explanation requests might exist too in order to provoke explanations from the other partner.

1 Introduction

Design as a creative process strongly depends on effective ideation and teamwork. Ideation has been described "as a matter of generating, developing and communicating ideas" ([27]). During ideation, divergent and convergent phases alternate to explore many possible ideas and select promising ones for further development. During the ideation process, designers seek inspiration from sources such as random encounters, browsing previous ideas, or teamwork. In particular, diverse teams have proven to yield better ideas by bringing together complementary perspectives and skills [3].

Ideation can be increasingly supported by software since advancements in artificial intelligence (AI) are leading to systems that can participate in ideation. This is a promising development because AI systems can provide new capabilities, such as processing extensive amounts of inspirational material in the background. With these capabilities, AI systems can complement human design teams in ways that other human collaborators cannot. This extends to so-called creativity support tools, i.e. software or user interfaces that aid creative expression or creative thinking [22]. These are a growing topic in Human-Computer Interaction and increasingly many of

them are co-creative [15]. A co-creative tool contributes to a shared piece of creative work together with the user [9]. These works are a promising development leading to systems that can directly work together with designers in ideation. They inspire by offering suggestions or contributing directly to a shared piece of work. Some examples of such systems suggest phrases for poetry [42] or additions to 3D models [6], provide inspirational images related to the creative discussion of a design team [55], or draw together with a user on a shared canvas [10]. Depending on the nature of the contribution of the system this can be regarded as a means to browse previous ideas, or provoking random encounters.

To maximize the advantage of the collaboration between human designers and AI systems it is worth studying the definitions and requirements of collaboration between humans. Collaboration has been defined as the "process through which parties who see different aspects of a problem can constructively explore their differences and search for solutions that go beyond their own limited vision of what is possible" ([18]). This definition highlights the advantage of bringing together different perspectives and skill sets. It highlights collaboration as an opportunity to add different perspectives as a source of inspiration for more profound solutions. Furthermore, other authors have included the need for shared objectives and strategies in their definitions of collaboration [8]. This emphasizes the necessity of finding these shared objectives. This works differently depending on the nature of the collaborators. For this thesis, the collaboration of small groups is most relevant.

Small groups often perform shared cognitive acts, such as learning or making decisions together. These acts can be attributed to the group rather than to any of the individual group members. This is known in research as group cognition [53]. In this way, collaborative work or learning can exceed the intersection and even the sum of the individual members' own cognition [3]. Group cognition is enabled through conversational grounding and theory of mind [31]. Conversational grounding the ability to reach a common understanding through communication, while theory of mind is the ability to understand one's own and other peoples' mental states. Both strongly depend on explanations: to understand each other's mental states and to identify or form common objectives, the members of a collaborating group need to externalize their reasoning. During individual work or learning, cognitive processes naturally stay hidden. For collaboration, group members need to educate the group about their reasoning to enable the group to take part in the collaborative process [53]. This understanding of each other's reasoning enables judgment of whether the other can be trusted. As well, it makes possible utilizing differences e.g. not only the contribution of another collaborator but also the reasons behind it may provide a potentially inspiring insight. Furthermore, the process of educating others helps also the explaining individual to reflect on their own assumptions and, if necessary, refine them [36]. In addition, for ideation in particular, verbalization has been identified as one of the main sources for inspiration [27]. These factors strongly motivate explanations as a means to make reasoning and perspectives transparent for oneself as well as others.

State-of-the-art creativity support tools with AI contributing during the creative process are progressing quickly in the way they inspire through providing external

stimuli or suggestions. Nevertheless, most of these tools do not include explanations. Only a few provide some information about how the algorithm created its output [4, 28, 43], and to the best of my knowledge, there are no ideation support systems that allow their users to explain their contributions. The interaction is limited to steering actions or the system utilizes observations as input. This lack of reciprocity of explanations can lead to a loss of some advantages that collaboration may bring, e.g. they might fail to increase understanding, trust, reflection, alignment of goals, and inspiration through reasons behind contributions.

To close this gap, this thesis introduces the novel concept of reciprocal explanations for the human-AI partnership in ideation. The concept was developed as an interaction technique for co-creative ideation where a human user, as well as an AI system, contribute to a shared piece of work. Reciprocal explanations are inspired by insights from group cognition. In human teams, a team member who does not understand the idea of another would ask for an explanation. The other team member has to provide such an explanation in order to convince that their idea is worth exploring. However, if they have not yet reflected on the reasons behind their idea, they would need to do so in order to produce one. In this way, the explanations are beneficial for the receiver as well as the sender. To facilitate the ideation of the human designer they should be enabled to provide and receive explanations. Hence, reciprocal explanations combine explanations in two directions. One direction is represented by system explanations. These are provided by the AI system and communicate the reasons for the contributions of the system to the designer. The other direction is represented by user explanations provided by the designer about the reasons for their contributions. There is also the option of prompts for explanations for a specific contribution, which are called explanation requests. The goal of reciprocal explanations is to create a sense of partnership by making explanations more equal. Furthermore, both directions provide insights about the ideation process by referring to contributions to the shared work. This is potentially inspiring for the user for both directions of explanations.

A detailed description of reciprocal explanations is provided in chapter 3. Before that, in chapter 2, the theoretical background and related research are discussed. In these previous works, various explanation techniques were contributed, especially explanations provided by algorithms have received significant research attention across disciplines [1]. Yet, none of them explore reciprocity. Furthermore, the effects of user explanations on the user have not been researched as well and the effect of system explanations, particularly during ideation, on the joint work process of AI systems and humans. Therefore, after defining reciprocal explanations, this thesis continues with the design and integration of a reciprocal explanation feature for a design ideation tool (chapter 4) and evaluation of the feature in a user study with 16 professional designers (chapter 5 and 6). The results are discussed and design implications are drawn from it in chapter 7. The thesis finishes with open research problems around reciprocal explanations for future work 8, and a conclusion 9.

2 Background and Related Work

This thesis is informed by previous theories and studies. In this section, I will highlight the literature providing the basis for this thesis, starting with definitions, followed by theoretical background and related work.

2.1 AI, AI Systems and Agents

This thesis regularly refers to AI (artificial intelligence). AI is defined by Nilsson [40, Ch. 1.1] as the field concerned with understanding intelligent behavior with the goal of building intelligent artifacts. Intelligent behavior is then further defined as involving "perception, reasoning, learning, communicating, and acting in complex environments". Nilsson refers to the artifacts build in AI as AI systems. Due to their definition AI systems describe a broad category of artifacts that behave intelligently. This broad understanding of AI systems is desired in this thesis because the concept of reciprocal explanations is independent of what exactly generates the intelligent behavior. However, the AI systems of relevance in this thesis are those that can be embedded in creativity support tools (CST) to make them co-creative – that is contributing to a shared piece of work with the user like a partner or assistant [9]. Hence, AI systems in this thesis mainly refer to the part of a software that provides the required intelligence to contribute to the shared piece of work.

Nilsson also uses the expression "agent" in place of AI systems [40]. I also use the term agent. However, I employ a definition of agents and agency that can be applied to AI systems as well as humans. Engen et al. [14] define agency as "capacity to perform activities in a particular environment in line with a set of goals/objectives that influence and shape the extent and nature of their participation." They further scope agency with three factors.

1. *Activities:* Which activities may the actor perform? Are they able to perform many different activities or are they limited in the activities, e.g. through access rights in an application?

2. *Nature of the Activities:* How diverse, free, creative, unpredictable can the actor be in their activities? Are the outcomes of the actions predefined (closed action) or can they be diverse (open action)?

3. *Interaction with other actors:* How big is the influence the actor can have on the other actors? The influence depends on their ability to communicate with these other actors.

According to these factors, actors may exhibit agency to different degrees. An actor that possesses some level of agency can be referred to as an agent. This theory of agency enables studying the effect of reciprocal explanations on the agency of the AI system and designers involved in ideation processes.

2.2 Mixed-Initiative Systems

Already for decades, there have been considerations on how humans and computers could collaborate effectively. J.C.R. Licklider [33] already wrote in 1960 about a symbiotic partnership between computers and humans and noted that together they could accomplish intellectual work more efficiently than humans or computers alone. In this section, I discuss especially those concepts that are especially relevant for the collaboration between humans and AI and using explanations.

Within AI-Human collaboration, the concept of mixed-initiative systems is relevant. The most common definitions of mixed-initiative interaction in Human-Computer Interaction emphasize that several agents (human and computer) solve a task jointly with contributions based on their knowledge and capabilities [25, 20]. E. Horvitz [23] identified 12 principles for mixed-initiative interaction. Some of them inform reciprocal interaction directly or indirectly. They suggest, that in case of uncertainty about the intentions of the user, the system should be able to clarify them in a dialog with a user and that there should be a quick and intuitive way to reject its services.

2.2.1 Mixed-Initiative Systems and Trust

Lieberman [34] emphasizes the importance of explanations in mixed-initiative human-AI collaboration to make users understand the AI system and consequently gain trust. Furthermore, transparency, reciprocity, and collaborative use of resources have been identified as important aspects for symbiotic interaction [24] as well as trust [14]. Transparency, meaning that the workings of an algorithm are accessible, is another aspect motivating explanations since explanations can be used to make the working of an AI system visible. Trust depends on the level of perceived ability, integrity, and benevolence [37]. One way to influence the perception of these factors could be to explain the principles behind the AI system.

2.3 Group Cognition in Ideation

The dynamics of human teams can also inform the collaboration of a human and AI system. Group Cognition studies the phenomena that small groups are capable of performing cognitive acts. From group cognition, Koch and Oulasvirta [31] reviewed how theory of mind and conversational grounding could enable AI systems to become more collaborative. Being able to form a theory of mind i.e. being able to anticipate the mental state of somebody else and oneself is important to predict a collaborator's next actions and adjust one's own goals and actions [31]. Conversational grounding is the process of reaching mutual knowledge, believes, and assumptions through conversation [31]. Both of these requirements for effective group work motivate reciprocal explanations as it strengthens the necessary exchange of information.

It has been found in a study with professional and student designers that spoken and written verbalization was the most used tool for ideation in practice and the tool that had the strongest link to moments of sudden insight ("Aha") [27]. The authors argue that this attributes to the social and collaborative nature of design and

the human preference for using language as the primary means of communication. The conversations during collaborative ideation are structured in so-called CI-loops (collaborative ideation loops) which are reoccurring patterns starting with *naming* an element of the design or stating a *constrain* such as a project budget to open a *negotiation* which gets concluded by a design *decision* or *moving action* i.e. a change of the representation of the design [11, 12]. Verbalization also plays a role in the communication of design. Designers get design briefs which are formulated verbally and need to be translated into visuals. Visuals are then discussed verbally throughout the design process among designers as well as between designers and clients [54]. However, many designers perceive finding verbal descriptions for their designs difficult [54]. These findings encourage us to develop reciprocal explanations in a textual form supporting design ideation by providing written descriptions.

2.4 Mood Board Design

This thesis studies reciprocal explanations on the example of a tool for mood board design. Creating mood boards is a common method in design practice with the purpose to ideate and frame a design project in the early stage [35]. At this stage, there exists only a vague idea of the purpose of the future design often as a design brief i.e. a textual description of how a client envisions the final product. The mood board usually consists of a collection of images, colors, and other materials that are related to the design brief [35]. These materials are brought together in a virtual or physical collage. The search and exploration of the mood board content help designers translate and abstract the textual idea into visuals and materials that convey the overall feel of the later design. The mood board is further used to communicate about these visual ideas with a client.

2.5 Explanations and Algorithms

Explanations are a wide and active research topic in the field of AI and Machine Learning, which is the field concerned with algorithms that learn from experience, thus automatically [38]. Most research concerns the implementation and effect of explanations of the output of algorithms. Apart from that, some algorithms can learn from user explanations.

2.5.1 Explainable AI

In common terms, explanations are mostly thought of as a specification of a cause, reason, or intention behind an action, event, or state [50]. They can be seen as a way to provide the information that is needed to make a decision [47]. What exactly constitutes a satisfying explanation depends on the context including the goals of the sender and receiver of the explanation [50]. For the case of explanations given by an AI system, the goals can be grouped into five categories [47, 50, 51]:

- Transparency: Telling how the system arrived at the output,

- Justification: The answer to the question of why the output is good,
- Relevance: For a system that asks for user input, this explanation goal is to justify why the question is relevant for the problem,
- Conceptualization: Providing definitions to reach a common understanding of the vocabulary of the system,
- Learning: Teaching the user about the domain

In the field of machine learning, explanations are mostly researched as a way to provide evidence in textual or visual form for how the input and the model's prediction are related [44]. There are two evaluation criteria in this scenario with the first one being interpretability and the second one completeness [16]. Interpretable explanations are given in such a way that a human user can understand them while complete explanations are fully accurate and reveal every step taken to reach a prediction [16]. These internal steps are usually difficult to make sense of for non-experts, in the cases of "black-box" models even for experts. Ideally, an explanation would present the complete reasoning of the algorithm in an interpretable way. However, this is challenging. Therefore, explanations often trade-off completeness against interpretability. A review paper on explanations in machine learning [16] identified three approaches: (1) Emulating the processing of the data to connect the input and output and produce a justification, (2) using internal representations to facilitate an understanding of intermediate steps that led from the input to the output, and (3) explanation producing algorithms that explain themselves [16]. All of these approaches have in common that they primarily focus on transparency and justification or a combination of both as goals. It is relevant to note that, in machine learning, transparency serves not only the end-users of a model but also helps developers evaluate it [56]. In this thesis, transparency and justification also provide a major motivation for explanations. However, goals such as conceptualization and learning, which are often ignored in Machine Learning, also play a role in this use case.

Machine learning is not the only research area investigating explainability. Abdul et al. [1] conducted a brought literature review across fields to identify which research communities explored explainability with which goals and results and how they are connected. Interpretable machine learning stayed, in fact, relatively disconnected from other areas including earlier AI topics such as Recommender Systems and Case-Based Reasoning.

Chandrasekaran et al. [5] conducted the, to my knowledge, only user study testing the ability of people to predict the output from a machine learning algorithm ("theory of AI mind") and if information about the internal states of the algorithm helps. It could be shown that they could learn from examples what an AI system would answer and when it fails. In this study, revealing internal states of the algorithm, such as confidence and attention maps highlighting how relevant each part of the input was for the output, did not help participants predicting the output. The authors of this study hypothesize that the reason was that the study participants

overfitted their predictions. The study encourages to develop and utilize different explanation methods that tailor better to the knowledge of users without machine learning expertise if the goal is to make an AI system more predictable to the user. Based on these results, I consider an additional translation step in my implementation of reciprocal explanations to abstract more from the technical features to avoid such overfitting effects.

2.5.2 Interactive Machine Learning

There are various approaches to machine learning algorithms that work interactively with a *human-in-the-loop*, meaning that they learn with the help of a human user [46]. These are often targeted at users who are not experts in machine learning and provide an interface that lets users generate data or manipulate a model directly [13]. The interface then typically provides feedback about the effect of the changes so that the user can iteratively improve the model. Related to these are approaches called active learning, which pick data that is promising for the learning outcome of the algorithm [49]. Combined human-in-the-loop and active learning approaches, then, let human users teach the algorithm by providing that data. These user inputs can be regarded as a restricted form of user explanation. Amershi et al. [2] studied three different interactive machine learning approaches where human users teach an algorithm and found that humans dislike being "used as oracle" and in some circumstances are also surprisingly poor teachers for algorithms e.g. providing too positive feedback which causes the reinforcement of wrong behavior. The anticipated reason for this is that "users want to allow future *right* behavior rather than give feedback to previous behavior." A more preferred way to teach the algorithm was to showcase how it would be correct. Furthermore, transparency was valued and helped to teach better. These results inform reciprocal explanation even though the focus is not primarily on teaching. However, if showcasing by examples is a motivating way to interact with an AI system, this could be applicable in explanations for user contributions.

2.5.3 Learning from Verbal Explanations

There are also some techniques to train machine learning algorithms through more natural explanations from humans. For instance, explanations in natural language have been used to train a classifier [19], teach an AI system play a video game (Mario Bros) [32] or concepts and classification in parallel for the use case of emails [52]. These are interesting recent contributions for algorithms that can utilize verbal explanations to improve the performance of an algorithm. They show that there is a research interest in utilizing user explanations for AI systems.

2.6 Co-creative CST

Creativity support tools (CST) are tools that help people in creative thinking and expressing themselves creatively [22]. In the computer domain, usually, creativity support tools are software that is used for creating digital artifacts or facilitate a part

of the process of creating an artifact [7]. Creativity support tools are developed for many use cases including but not limited to photography, music, writing, architecture, or design. Ideation is a creative process and a relevant part of design. Tools that support ideation can clearly be classified as creativity support tools. A survey found that creativity support tools represent a growing number of contributions in the field of Human-Computer Interaction and that is a strong focus on tools that support collaboration [15]. In most tools, collaboration support is focused on solely human collaboration. However, co-creative AI systems that contribute like a partner or assistant to the creative artifact are also an increasing trend. These tools could potentially benefit from the reciprocal explanation technique.

2.6.1 Interactive Behavior of Co-Creative CST

In this section, I focus on those co-creative creativity support tools where at least one, either the human or the system explains their contributions or prompts explanations during the creative process i.e. after the setup phase.

Some tools provide context or references to sources as an explanation of how its output relates to the input. This is the case for a tool introduced by Baumer et al. [4] which is meant to facilitate critical and creative thinking by displaying metaphors. These metaphors are computed by comparing the context of words in different domains. The explanations consist of the context of the words and support the critical thinking process. The tool CombinFormation [28] for information discovery suggests content for a collage of information. Further information about how they were found and a reference to the source can be accessed in a menu. Another tool creates suggestions for additions to 3D models also by comparing the context [6]. In this case, 3D models are compared to the work of the user to examine if they have parts that could be added. The suggested additions are then displayed along with their original context as an explanation where they come from and how they could match. These explanations are interesting in that respect that they connect to how the system found them, hence provide transparency, and also contain some information that could inspire.

ReQUEST [45] is an approach to help write stories by asking questions that potential readers might have. The idea is not to contribute directly to the story but to stimulate reflection while it is written. The algorithm detects when there is something that might appear illogical and points that out with questions about motives and consequences. The user answers the questions by simply editing the story further. With the explanation requests in reciprocal explanations, I also aim at provoking reflection. Another system pointing out inconsistencies in a story was designed by Samuel et al. [48]. It presents almost consistent options for continuing the story along with an explanation of why they are not consistent. There were three aims of presenting these options (1) educate the user that the option is inconsistent in case they were planning to continue with it, (2) making transparent what the state of the system is, (3) inspire a change of the story that would make the option possible. Transparency and inspiration are also among the motivations of reciprocal explanations. However, both systems were not evaluated for their true effect on the

users with a sufficient amount of participants. The evaluation of the effect on users remains a gap.

Duet Draw [41], is a tool with an AI agent that draws cooperatively with the user on a shared canvas. It can give instructions and explain its intentions. Varying levels of detail in the explanations were tested with users with the result, that users preferred the more detailed explanations.

In summary, it can be said that reciprocity is still lacking across explanation techniques. Most approaches have been developed to explain the reasoning of AI systems. Those explanations that can be received from users typically serve the needs of the system, e.g. for training. Reciprocal explanations aim to close this gap by combining explanations in both directions and target them at improving the joint ideation process, rather than the AI system. I want to achieve that by linking each explanation to a contribution either from a user or the AI system.

3 Reciprocal Explanations

This chapter introduces reciprocal explanations as a novel technique for explanations. It will discuss the motivation for adopting reciprocal explanations and introduce the research question and hypothesis of this thesis.

Reciprocal explanations are a technique for explanations in mixed-initiative interfaces. In these interfaces, several agents, human and computer, take initiative and contribute to a task according to their ability and skills. With reciprocal explanations, both humans and computers, provide reasons for their contributions.

Two directions of explanations are necessary for reciprocal explanations. One direction entails explanations provided by the system. These *system explanations* inform the user about the reasons for its actions. System explanations can take the forms in previous work on explainable artificial intelligence. The other direction entails a way for the user to provide further information about their contributions to the system. I refer to these explanations as *user explanations*.

There needs to be an interface for both directions of explanations. The initiative for either direction of explanation could come from either agent. Each one could provide an explanation along with their contribution. Alternatively, there could be an option to prompt explanations by issuing an *explanation request*. Which initiatives to implement needs to be considered for each use-case of reciprocal explanation. Explanation requests only make sense if the agent does not already explain all contributions by default. One more consideration is the frequency of explanations and explanation requests, e.g., too frequent requests potentially become disruptive, while too infrequent ones might lead to a loss of reciprocity.

I developed the technique of reciprocal explanations with co-creative tools for design ideation in mind and focus on this use case in this thesis. Nevertheless, co-creative tools can be regarded as a sub-category of mixed-initiative systems that specialize in creative tasks. Even though I evaluate the concept only for this specialized use case, it is likely transferable to other mixed-initiative interfaces.

3.1 Motivation

The goal of reciprocal explanations is to facilitate the co-creative ideation of a human user and an AI system. The purpose of the AI system during the ideation is to inspire the user as a work partner. This requires the ability to provide diverse contributions as well as appropriate communication [31]. Reciprocal explanations are designed to improve the communication. The following section details the hypothesized benefits.

First of all, reciprocal explanations might emphasize the partnership of the user and AI system by creating a more even interaction with the sense of an AI system that has agency. A system with an agency has its own agenda including own ideas, options for actions and influence [14], which in turn is a requirement to add value as a partner with its own ideas[31]. Reciprocal explanations might communicate this value to the user. The ability to explain and to request and receive explanations alone are additional activities that the AI system can perform, and with an increased scope of activities, the agency increases. Furthermore, if the AI system has its own

ideas and reasoning, this can be communicated via explanations, and increase the awareness of the user that there is more to the AI system than a simple amplification of what they are already accomplishing. Additionally, an agent can use explanation requests to point out that it has a perception of the user's contributions, also hinting toward its own agenda.

A similar yet slightly different goal is to communicate the abilities of the AI system, meaning that apart from showing that it has its own agenda, reciprocal explanations can also communicate that it pays attention to relevant features. In one direction explanations may justify that the suggestions are good, in the other they can communicate that it is able to judge other contributions. However, this depends on the concrete design of the AI system and feature, e.g. explanation requests could also appear as if the system depended on input from the user or asks because it is particularly uncertain. If the reciprocal explanations can communicate the competence of the AI system, this is likely to also have a positive effect on the trust since ability is one of the important factors for this [37].

A further motivation is to simulate parts of the grounding happening during creative collaboration where team members discuss and explain their knowledge and ideas creating a common understanding and alignment of goals during the design process. This process should be reciprocal and continuous throughout the design process [31]. Ideally, the designer can evaluate with the help of system explanations if they agree or disagree and consider aligning with the objectives expressed through it. In other words, the designer might find inspiration for a change of strategy not only in the suggestions but also in the explanations for them. At the same time, the designer should be able to justify their own decisions to the system e.g. why they selected an image. The system can use this information to align its strategy to the strategy of the user. Strategies can change over time during the ideation process. Transitioning from a divergent to convergent phase or vice versa can occur or a serendipitous inspiration can alter the direction [17]. Hence it makes sense to continue asking for user explanations from time to time.

Commonly, machine learning algorithms prompt feedback or ask for data input which can be used for training the algorithm. In such a training scenario the user functions as a teacher but does not need to reflect their own decisions or think about why they made a choice. However, in ideation, such awareness is potentially useful for continuing the work. The inspiration for adding reciprocity to aid reflection comes from human collaboration. If one of the collaborators does not understand the reason behind the other's action, they would ask. The other than needs to provide a reason. If they are not aware of these reasons yet, answering would require them to start reflecting. After receiving an answer, the first then has the possibility to align their work but can choose not to if it is against their interest. Reflection can also lead to recognizing inconsistencies in the own reasoning and thus in a refinement of this reason. Implementing reciprocal explanations could lead to an improvement in these processes of understanding work partners' as well as one's own reasoning.

3.2 Research Question

There are some open questions that need to be answered to make the most out of reciprocal explanation in an ideation tool for designers. Even though previous researchers have developed co-creative systems only a few of them include a sort of explanation [4, 28, 43]. None of them include two directions of explanations, leaving the interaction on uneven terms. Furthermore, there is a lack of understanding of the effect of user explanations during the interaction with the system. The following question remains to be investigated.

How can reciprocal explanations provided by the computer and the designer facilitate their co-creative work?

The motivation of the design of reciprocal explanation is improved reflection of the designer on the design process and their ideas, improved understanding and alignment of strategies through a sort of conversational grounding between the designer and AI system. Furthermore, I hope to emphasize the existence and inspiration through the AI system's independent ideas. I would like to investigate which of these benefits are achievable.

This question can be broken down into two sub-questions making the research question more approachable for evaluation:

1. Which benefits can be reached through system explanations and user explanations separately?
2. Is there an additional benefit achievable through the combination of the two, which is more than the sum of two?

3.3 Hypothesis

Previous work inspired and motivated the concept of reciprocal explanations. It leads to the following hypotheses which summarize the anticipated benefits described in the subsections above 3.1.

H1: Reciprocal explanations help improve the theory of mind, meaning that designers better understand the reasoning of the AI system and themselves through reflection.

H2: Reciprocal explanations aid further-reaching exploration by reminding the designer to think about different features and strategies.

H3: Reciprocal explanations aids the alignment of goals and strategies between the AI system and the designer.

H4: Reciprocal explanations can communicate the agency and agenda of the AI system to the designer.

A complete answer to all of these hypotheses is out of the scope of this thesis. However, I aim to provide at least partial answers to all of them.

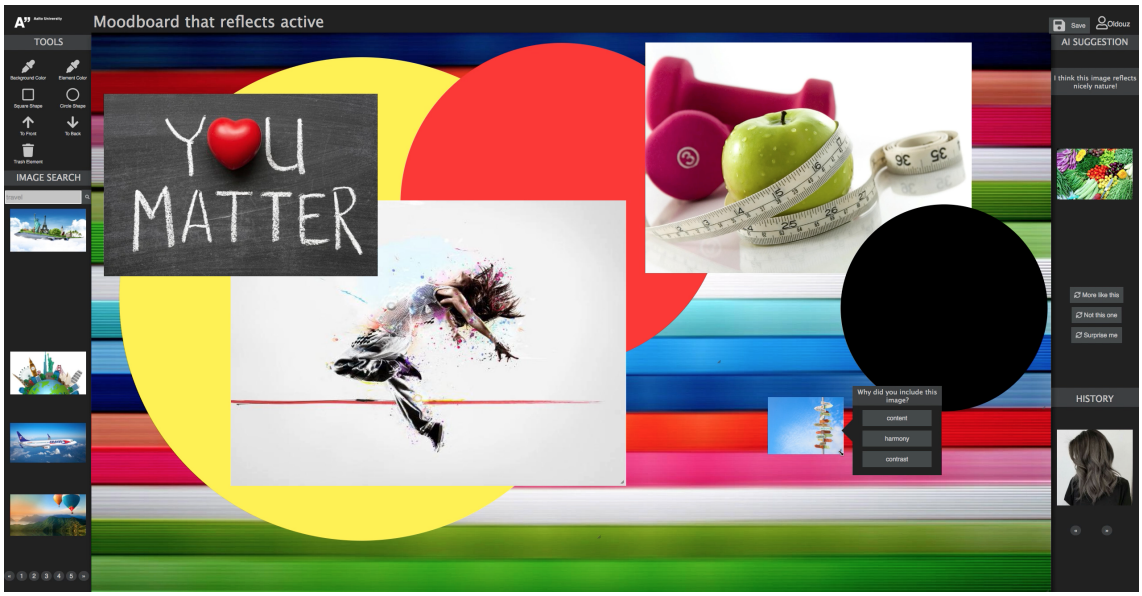

Figure 2: The interface of the May AI tool [30]. The largest space in the middle has a canvas to hold the mood board. On the left side, there is a panel with tools to manipulate the mood board and an image search. On the right side, the AI makes suggestions.

4 Design and Integration

This chapter deals with the design and implementation of the reciprocal explanation feature for testing purposes. It is integrated into a tool for mood board design called May AI described in detail in [30]. Mood board design represents a popular method for ideation, which entails creating a visually stimulating collection of inspirational images and other materials [35]. In this chapter, I only present as much detail of the May AI tool as necessary to understand my implementation.

In summary, the May-AI interface is used as follows. When a designer decides to create a mood board, they can log in with a user name of their choice. At the login, they also provide a title for the mood board which is then used to create a new empty mood board. To the left of the mood board, the designer can find an image search interface where they can query images from the web[1], drag them onto the mood board, and change their size as needed. Above the search, there are tools to manipulate the mood board, such as changing the background color, adding shapes, or changing the arrangement of images. On the right of the mood board, there is a panel interfacing with the AI. Here, the AI displays image suggestions. The designer can choose to use them by dragging them on the mood board. If they do not consider the image suitable, they can request more similar or different images using buttons, or ignore them and continue adding images from the left. In this case, the AI will display a different suggestion every time the designer adds a new image to the mood board.

[1] https://duckduckgo.com

Feature	Value Range	Description
Dominant Hue	[0, 360]	The hue of the dominant color is given as a degree on a color wheel with red at value 0°, green at value 120° and blue at value 240° and additively mixed colors between these values.
Dominant Saturation	[0, 1]	The saturation describes how much of the hue is present in the dominant color. The closer the value to 1 the more vibrant the color. 0 describes a white, black, or gray shade.
Dominant Lightness	[0, 1]	The lightness describes the amount of black or white mixed into the color. Values below 0.5 are darker with a maximum amount of black at 0. Values above 0.5 are lighter with a maximum amount of white at 1.
Color Contrast	[0, 180]	The color contrast is the difference between the hues of the two most dominant colors. It is given as the difference of the degrees in the color wheel. It is 0°if the hue is the same and 180°if the hue is located on the opposite side of the color wheel.
Image Orientation	{horizontal, vertical}	Relation of the height and width of images. A horizontal image has a longer width than height. For a vertical image, this relation is vice versa. The orientation of the mood board is the prevalent orientation of the images on it.

Table 1: The features used to describe the design space in May AI. The dominant color is the mean of the biggest cluster of color in an image.

The first step to produce a suggestion is to analyze the visual features of the mood board. This is done every time after the designer added a new picture to it. The features are based on dynamic color clustering. The largest clusters are used to determine the dominant hue, saturation, and lightness, as well as the color contrast. Additionally, image orientation is used. More details about the features can be seen in table 1. These features were selected based on observable differences between mood boards. A vector of these features is passed to a cooperative contextual bandit algorithm as a context.

The AI system is implemented using cooperative contextual bandits, which is a recommender system with the ability to make suggestions based on given features of a context, learning the preferences of a user, and to balance exploring and exploiting. Given a multidimensional design space with the features as dimensions, the system tests suggesting vectors from this space. From these reactions of the designer to the suggestions, the system can learn their preferences. The assumption is that, as long as the strategy of the designer has not changed, similar suggestions close to each other in the design space would result in similar reactions. Apart from the learned preferences, it can also consider the context to decide what to suggest. The system can balance between exploitation and exploration strategies. An exploitation strategy makes suggestions close to the given context and the previously accepted suggestions. An exploration strategy tries suggestions from areas in the design space that are more different from the context and that haven't been suggested to the designer yet. A mixture of these strategies is needed, considering previous choices of the designer, the current context, and bringing up examples from new but related areas. To realize such behavior, the design space is spliced into so-called strategies. Each strategy is represented by a so-called *strategy agent*. The strategy agent representing the features of the mood board is called for a feature vector. To exploit, this strategy agent returns a vector from its own strategy. For exploration is passes the decision to one of the neighboring strategy agents. The probability of the bandits exploring or exploiting can be set with a cost parameter for exploration c_m^n. This parameter can be changed during runtime. In May AI, the designers can influence it, e.g. with the steering buttons under the suggestions.

The next step of the algorithm is to find an image that matches the chosen feature vector and the topic of the mood board. The system is equipped with a database, storing readily computed feature vectors for previously used images along with the search terms that were used to find the images. The system will first try to lookup an image from the database fitting the desired feature vector. The system is also keeping an association list containing words queried from an association API using the mood board topic and recent search terms of the designer. The associations are used as search words for the images to ensure that the images are relevant to the topic of the mood board but also divergent from the designer's by adding the associated words. If no image from the database matches the associations and features by a certain tolerance, the system will start searching for images using the online search engine

Duckduckgo[2] and analyze the features until it finds an image matching or until the designer moves on changing the mood board.

4.1 Reciprocal Explanations

I designed and integrated the first version of reciprocal explanation into May AI. The May-AI tool makes suggestions considering key words, exploration, and exploitation strategies. The design of reciprocal explanations for this tool takes these considerations into account as *content, contrast*, and *harmony*. By considering associated words to search for images, May AI finds images of relevant content. Harmony is ensured by using the current mood board as context and suggesting images employing an exploitation strategy. Contrast comes into play when the bandit system explores different areas of the design space. In this case, the AI brings up images that contrast the current mood board with some features.

I implemented reciprocal explanations around these three themes: "harmony", "contrast" and "content". Harmony and contrast are visual properties in this context. Colors or images as a whole can be close to each other resulting in a harmonious look or they can be distant in their visual appearance resulting in a more contrasted look. Both of these properties can communicate an abstract message hence they need to be considered and balanced. The third theme is content which is about the motives of images. These also have to be relevant to the mood board and convey a feeling, idea, or abstraction.

I use these three themes for the explanations in both directions of explanations to keep them consistent with each other and with the properties of the underlying AI system. This is relevant to make interpretable to the designers what the system pays attention to. Interpretable explanations for users who are not experts of AI require a translation into terms known by the user group [44]. The three themes can be regarded as translations of the strategies and usage of key words. Harmony is, furthermore expressed as *matching* and contrast as *complementing*.

The system explanation consists of a sentence with a reason why the image it is suggesting fits the mood board. The AI explains all of its suggestions comparing features of suggestions and context. This has been argued to be a good approach to make understandable justifications [21].

The AI system requests explanations for selected contributions from the user. The selection is made to keep the questions meaningful. For these explanation requests, a small pop-up window opens, pointing at the newest image on the mood board. In this window, it asks why the designer added the image. The designer may provide the user explanation by selecting one of the options for an answer. Details can be found in the following subsections.

4.2 System Explanation

.

[2] https://duckduckgo.com/

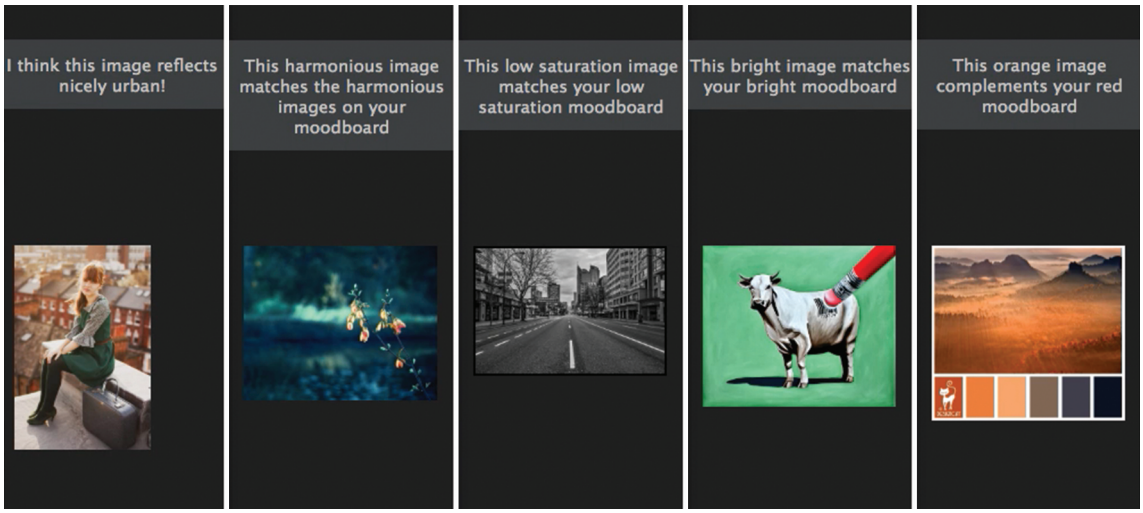

Figure 3: Each suggestion from the AI system is displayed together with an explanation either using a key word or relating the values of a visual feature in the suggested image and the current mood board. The examples in this figure explaining using the following properties from left to right: a *word* from the association list, *color contrast, saturation, lightness*, and *hue*.

System explanations are provided for every suggestion from the AI system. The reason for this is that it can be displayed unobtrusively. Hence, it is unlikely to become annoying even if it is continuously. One alternative is to show explanations only on request. This was not implemented to avoid complicating the access to explanation to an extend that requires learning how to request explanations and takes longer. A second alternative to always explaining is being selective on the system side about when to show an explanation. This approach is questionable because there is no clear rationale for making a selection and withholding explanations without good reason can be even considered unethical [16].

4.2.1 Structure and Appearance

Explanations are made in the form of a written sentence relating the features of the mood board to the features of the suggestion. For example, “This green image matches your green mood board” refers to the hue value green in the suggested image and the mood board and relates them by saying that they match i.e. both are green. More examples of explanation texts are shown in figure 3. Text is also the most common form of explanation between humans hence the most natural way to explain [27]. This is good, especially because the target group cannot be expected to have a background in machine learning and might not understand internal representations or the raw numbers of the features without translation into common terms. Translations are also able to abstract the internal representations which prevent overestimating the influence of individual parameters on the output of the AI system. Such overfitting effects can have a negative effect on understanding an AI system [5]

Yet it stays important to relate to the input as well as the output in the explanation. We cannot expect that the designer would evaluate the mood board in the exact same terms as the AI, hence we need to make it obvious in the explanations. Even if the explanation points out the value of a feature only in the output it might not be clear why this would be a matching choice without comparing to the corresponding value in the context.

The suggestions are based on the features of the mood board as a context. The bandit system chooses to present an image that either stays within the same strategy in the design space, selecting a set of features close to those of the context within a certain range. Alternatively, it chooses to diverge further with one of the features exploring a different neighboring strategy. In the first case, the selected image will match the mood board with all features. In the latter case, it potentially complements them. One of the purposes of exploration is to find types of images that are still missing on the mood board and hence complement what is on it. For this reason, the relation between the features can best be described as matching or complementing. I use these two descriptions in our sentences, resulting in a structure as

This red image matches your red mood board.

This orange image complements your red mood board.

Revealing the comparison might help designers develop a theory of how the system works. I designed the explanations in such a way that they can include the context, features that the system uses to find suitable images and search words for querying images. These factors are the basis for the system to find images. Over time an increasing amount of this information becomes visible to the designer. Hence, the explanations provide the opportunity to learn more about the reasoning of the AI system.

Knowing how the system reasons allows the designer to steer the AI system through their own behavior e.g. by altering the mood board. Such a strategy diminishes its agency less than e.g. taking over a commanding role to which the AI system can only react with limited responses. It would also decrease the bafflement about unpredicted output and hence increase the perceived integrity of the AI system. Higher perceived integrity would likely have a positive effect on trust [37].

I aim at inspiring the designers with additional information contained in the system explanation. The bandit system analyses images for their visual features, which are required to find a matching image. Revealing which features are visible in the current mood board and suggestion can bring the designer's awareness to these features. Especially, making them explicit in a text could be a source of inspiration to start paying attention to this feature or look for other images with similar features. If this happens, it could be interpreted as a form of value alignment between the system and the designer.

The visual appearance of the system explanations is designed to be unobtrusive i.e. integrated with colors and shapes similar to the rest of the interface so that it delivers additional information without drawing overly much attention away from

the actual suggestion. To improve readability the dark gray text is underlaid with a lighter gray background.

The timing of the system explanations is so that it always appears at the same time as the suggestion and stays as long as the suggestion is displayed in the panel. When the AI system starts searching for a new suggestion, the old suggestion with its explanation disappears.

4.2.2 Selection of the Feature for an Explanation

In order to keep the system explanations easily comprehensible, not all features are included in each explanation. The explanation would easily become cluttered and too long with several features. This would take up too much space and would distract. A selection also makes sense because not all features are always equally relevant and visible.

It is unlikely that to the designer all features always matter the same. E.g. in a mood board that contains only images with low saturation, saturation would be a relevant feature, either trying to contrast or match that. In this case, another feature might be less relevant. For instance, the hue shows less in images with low saturation, hence arguing with the matching or contrasting of the hue is unlikely to provide much value.

There might also be implications for the relevance depending on whether the AI chose an exploration or exploitation strategy. If it explores a new area of the design space, the feature that distinguishes the current and the explored strategy might be more relevant. However, it could also be the feature that keeps a strongly visible connection between the two is more convincing. When the system exploits the same design space, the argument of the explanations relies on similarity to the context which would count for all features but will have different visibility.

From the bandit system, the vector describing the suggested image is known. By comparing this vector to the context vector we know if the system chose an exploration or exploitation strategy. This is the basis for the explanation. If the strategy was exploitation all features are fairly similar. In the case of an exploration strategy, one of the features has a bigger difference.

The explained feature should be easily visible in the image and mood board, so that little effort required for the user to verify if the image and mood board really have the feature as mentioned. Hence, picking a feature that is easily visible is the priority. The visibility of a feature depends on the values of the other features. The hue is an obvious example of this. Per definition, the value of the saturation defines now much of the hue is present. Hence, the visibility of the hue depends on the value of the saturation.

The relevance of the feature also depends on the value of the feature itself and how descriptive this value makes the feature. In this case, this is less because of the visibility but because of how the feature characterizes the image. For instance, an image with a low or high lightness would be described as dark or light, but a medium lightness would typically not be pointed out.

My feature selection algorithm goes through the features one by one with a greedy approach i.e. it evaluates its visibility and descriptiveness based on thresholds and stops when it finds a feature that is promising in this respect. When it found a feature that is visible according to these thresholds it uses it for the explanation. Using practical testing I set the thresholds and ordered the feature checking so that explanations of all features are likely to occur over time. In cases that none of the visual features passes the relevance test, the explanation utilizes the word from the association list that was used to query the image. The following pseudo-code provides the general logic and order of feature selection

```
if hue is meaningful in suggestion
        and hue is meaningful in mood board:
    use hue for explanation

else if lightness is meaningful in suggestion
        and lightness is meaningful in mood board:
    use lightness for explanation

else if saturation is meaningful in suggestion
        and saturation is meaningful in mood board:
    use saturation for explanation

else if contrast is meaningful in suggestion
        and contrast is meaningful in mood board:
    use contrast for explanation

else:
    use image of query word for explanation
```

The following pseudo codes show the relevance checks with thresholds as used in my implementation. For the hue, it checks if the saturation is high enough (above 0.15) and the lightness within a range that makes it visible (between 0.15 and 0.9). Saturation and lightness both can have values between 0 and 1. Furthermore, the hue is excluded from explanations if either image or mood board has a strong contrast (more than 120). Contrasting images would likely be described as colorful or contrasting and not as dominantly of one color. Color contrast is given in as a difference of hues degrees between 0 and 180 (see figure 4 for the description of hues as degrees in a color wheel).

```
hue_meaningful(saturation, lightness, contrast):
    return saturation > 0.15
        and contrast < 120
        and 0.15 < lightness < 0.9
```

The lightness plays the biggest role if it is high or low. Black or white are added to the color for values below and above 0.5 respectively. Medium lightness (between 0.3

and 0.7) means that less black or white is added to the color, which makes lightness less descriptive.

```
lightness_meaningful(lightness):
    return lightness < 0.3 or lightness > 0.7
```

The saturation is most noticeable when it is particularly low (below 0.25 in our implementation), giving the image a greyer appearance. In addition, high saturation makes a color appear particularly vivid (0.75 in our implementation). This can also be worth mentioning in an explanation. Since saturation has an effect on the hue, it also becomes less relevant when high or low lightness levels interfere. However, our function does not check this, as brightness is tested before saturation and will be used for the explanation when it dominates.

```
saturation_meaningful(saturation):
    return saturation > 0.75 or saturation < 0.25
```

We use contrasts below 60°for explaining with the harmony of colors coming from the same third of the color wheel. Contrasts above 120°tell that the dominant colors have strongly differing hues. In this case, the explanation can point out a strong contrast.

```
contrast_meaningful(contrast):
    return contrast<=60 or contrast>= 120
```

4.2.3 Feature Translation

After selecting a feature, its value for the suggestion as well as the mood board needs to be translated into natural language wordings that can be inserted into the textual explanation. I implemented a translation function to translate hue, lightness, saturation, and contrast. Query words to not need translation because they are already in natural language. In the following, I will describe the implementations of feature translation.

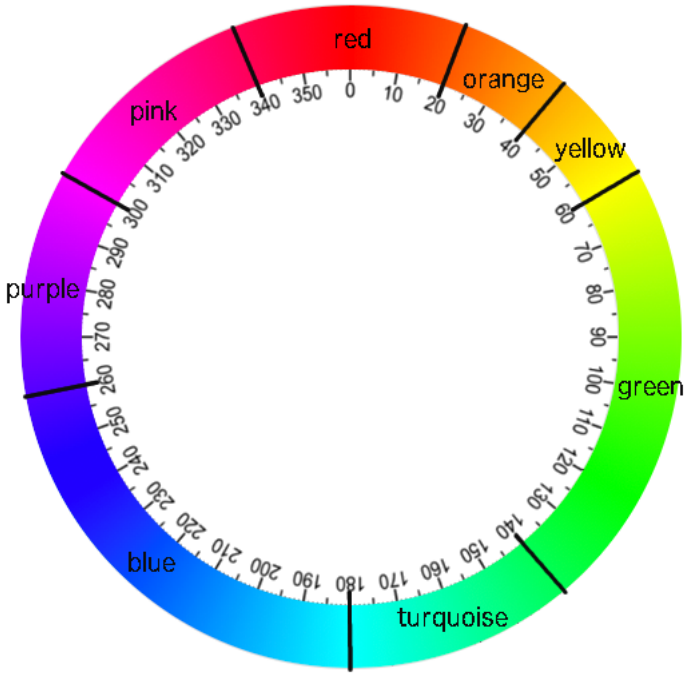

Figure 4: Partitioning of the HSL hues into the eight colors red, orange, yellow, green, turquoise, blue, purple and pink

Hue can be described as color names. Therefore the task is to translate a degree in the color wheel into a natural language color name. The hue is separated into six strategy agents in the underlying bandit system. Hence, the first approach was to use the names of the three primary colors in the RGB color space (red, green, blue) and the secondary colors (cyan, magenta, yellow) which each is the result of mixing two of the primary colors. However, aligning the color names with the separation of the bandit system was not practical because the bandits separate the colors exactly at the degree where the color matches the name best. For this reason colors similar to the pure primary and secondary colors, which could be described with the same name fall into two different strategies. Furthermore, the RGB color space and its color names are rather technical. For these reasons, we translate into color names based on the color wheel more commonly used in arts which are based on the primary colors red, blue and yellow, and the secondary colors purple, green, and orange. In the HSL color space, this results in uneven angles of color borders. Especially yellow and orange cover only small angles in the HSL color wheel while green and blue occupy large parts of it. To split these large parts into more precise slices, turquoise was inserted between green and blue and pink was inserted between purple and red. See figure 4 which shows the partitioning and translation of the HSL hues based on their degree values. Values for lightness below 0.3 are translated into the dark. Values above 0.7 are translated into light. An implementation for values in the middle is not needed because explanations will not use lightness for these values (see 4.2.3).

```
if bright_value < 0.3:
    return "dark"
else if bright_value > 0.7:
    return "bright"
```

It is difficult to find non-technical terms for saturation. Often highly saturated colors are described as bright. However, this could be confused with a high lightness. Other descriptions such as intense or vivid convey content that might not generalize well to all highly saturated images. For this reason, the translation stays close to the technical term. In particular, designers are knowledgable about saturation from working with color in other digital tools making an explanation containing this term appropriate. The translation function returns high saturation for values above 0.75 and low saturation for values below 0.25.

```
if sat_value > 0.75:
    return "high saturation"
else if sat_value < 0.25:
    return "low saturation"
else:
    return "medium saturation"
```

The contrast is based on the angle between two hues in the color wheel where distant colors are contrasting and closer ones are often described as harmonic [39] . Therefore, the translation function uses the terms harmonious for angles below 90°and color contrasting above.

```
if con_value <= 90:
    return "harmonious"
else:
    return "color contrasting"
```

4.2.4 Search Phrase Precision

Having natural language translations of the image features is useful to search images online more precisely. As mentioned above, the system queries and analyses images from an online image search engine, if there is no matching image in the database. The image analysis is time-consuming resulting in perceivable waiting times in the interface.

For this reason, it is necessary to find a trade-off between reactivity and how close the features of the image match to the features selected by the bandit system. To find images faster, the system can increase the tolerance for the difference between the image features to the features from the bandits. However, increasing the tolerance leads to suggestions that have a lesser link to the selection of the bandits. Furthermore, the bandit system cannot learn meaningful preferences from the user's decision if the image is poorly linked to its suggestion. The other extreme is to analyze images until one closely matches the desired features. This easily leads to processing times longer than it takes for the designer to change the mood board, which makes the search obsolete.

For this reason, I developed a way to extend the image search phrase with feature translations to narrow down the image search results to candidates closer to the desired features. Search phrases describing the desired feature vector as completely as

possible in a yet human-like way helps to find a matching image more likely on top of the search results, and hence faster. However, just concatenating the translations as above did not lead to effective search phrases. This was mostly due to the ambiguity of separate descriptions of saturation and lightness. For this reason, I developed a combined description of saturation and lightness. This description was used as a modulator phrase before the hue name e.g. pastel yellow. For finding images with large contrasts of more than 90°the name of a hue with the desired contrast was added after the first hue name. The search phrases have the following structure. The query word is the word from the association list, used as the original search term.

```
"<query word> <saturation-lightness-modulator> <primary hue name>
<contrasting hue name>"
```

This results in search phrases such as these examples:

```
"Vegetable dark red"
"Youth pastel green pink"
```

For any hue, there is a two-dimensional space defined by saturation and lightness, which my algorithm separates with easy to calculate borders into dark, pale light, pastel, gray, or no modulator phrase (see figure 6). The phrases are based on trials of image searches with various query words and hue names and subjective judgment of colors by several researchers from the lab.

```
# segments the color space into "dark" , "gray", "pastel", "pale light"
# or "" depending on the saturation and lightness value of the features

function get_modulated_color_description(saturation, lightness):
    if (0.25 < lightness < 0.6) and saturation > 0.5:
        return ""

    elif (saturation > 0.5 and lightness < 0.25)
            or (saturation < 0.5 and (saturation*2) < lightness ):
        return "dark"

    elif lightness > 0.9:
        return "pale light"
```

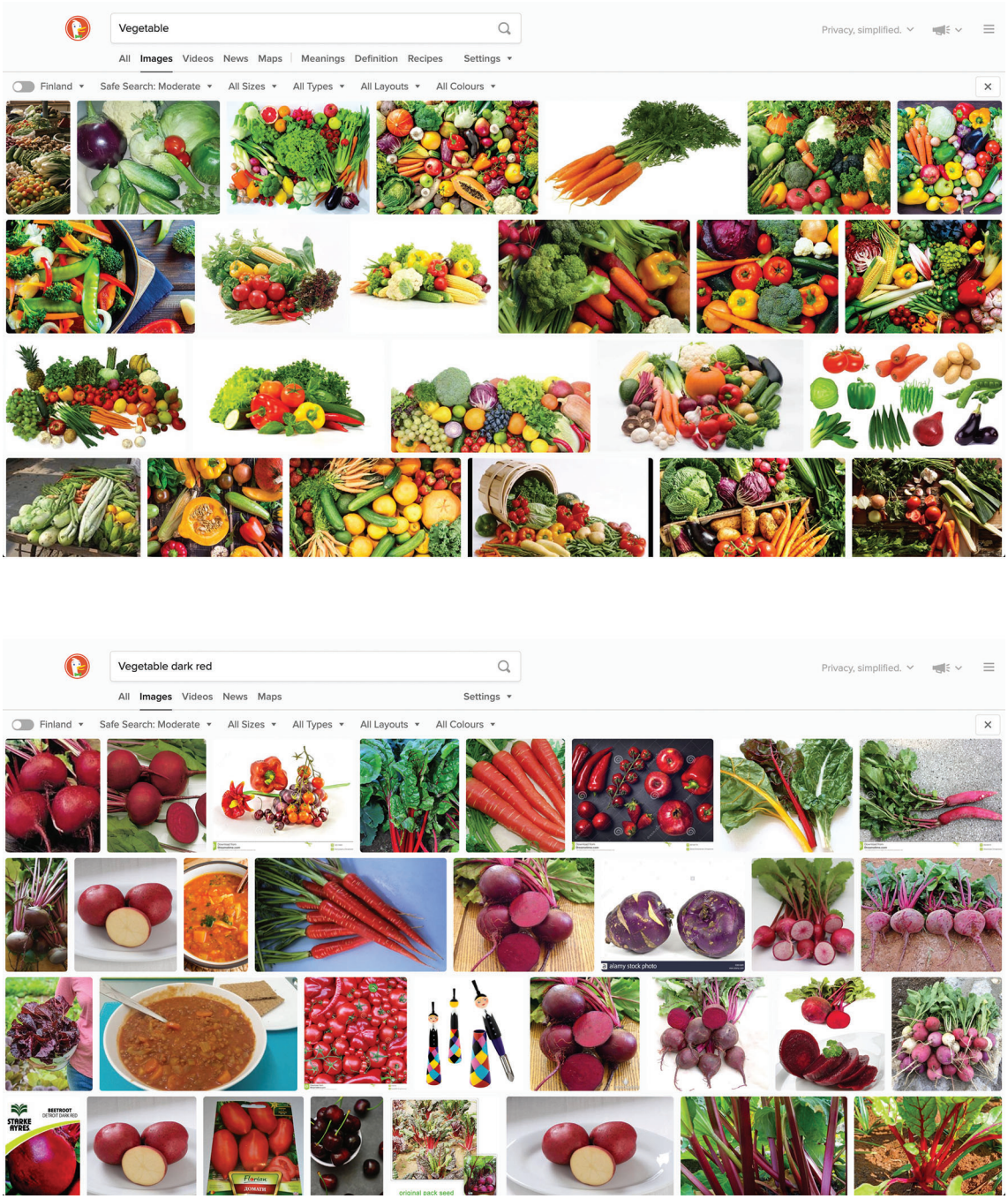

Figure 5: Examples of an unspecific search phrase using only a keyword (top) and a search phrase that describes the color value of the selected feature vector (bottom). Finding an image that satisfies the requirement of having a dark red tone as dominant color is more likely to succeed within the first search results with the optimized search phrase. (Screenshots from https://duckduckgo.com/, March 17, 2020)

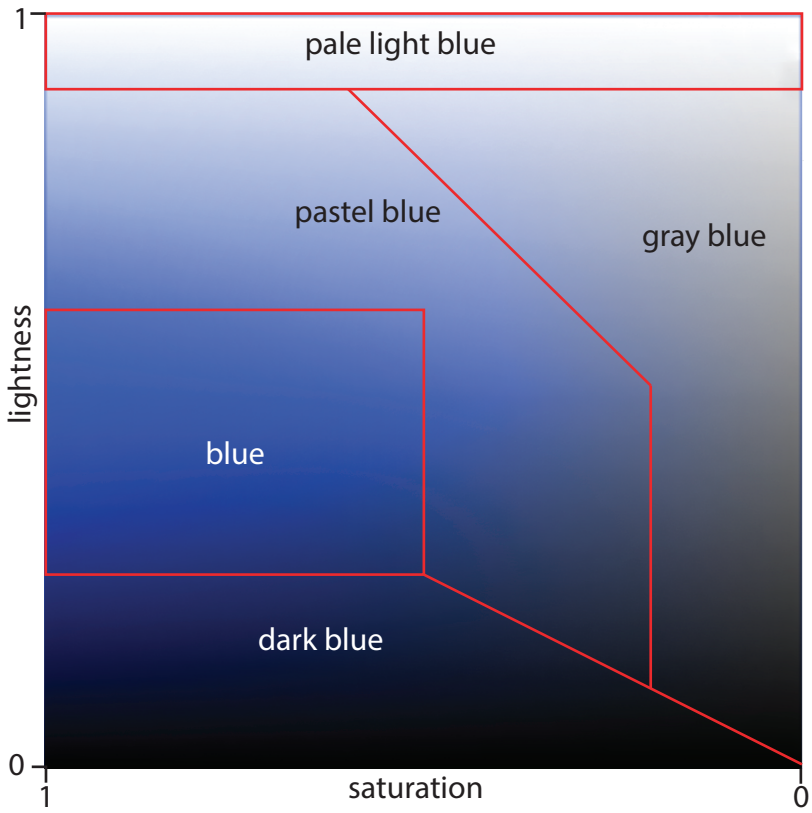

Figure 6: Partitioning of the two dimensional lightness and saturation space and the natural language descriptions pale light, pastel, gray and light that were used to optimize online search queries.

```
elif lightness > 0.5:
    if (-0.2+saturation) < lightness:
        return "pastel"
    else:
        return "gray"

else:
    if saturation > 0.2:
        return "pastel"
    else:
        return "gray"
```

With these descriptions, it was possible to reduce the tolerance for mismatched features significantly while at the same time reducing the search time for the majority of cases. In some other cases, search times continued to be longer than it takes to change the mood board and make the search obsolete. I suspect that they contain combinations of query words and features that are generally hard to satisfy. This could be the case as some query words might have strong associations with specific colors or some color combinations are rare.

4.3 User Explanation

In my implementation of user explanations, designers can provide explanations by answering a question from the AI system. The following subsections will detail on the design considerations and integration of user explanations into May AI.

4.3.1 Appearance

The user explanation is realized as a pop-up window in a visual style matching the rest of the interface. It pops up next to the most recently added picture on the mood board and has an arrow on the side which points at the image to make clear which image it refers to. It has a heading section containing a question about why the designer included the image. Below, in the body part of the pop-up, there are three buttons offering answer options for the designer. These options are *content, harmony* and *contrast*. These are chosen so that the AI can react to the answers by adapting its strategy.

4.3.2 Timing

To make explanation requests meaningful, and prevent possible annoyance by it, the AI needs to make a judgment of when it is appropriate to make an explanation request.

The explanation requests should not be disturbing or interrupting the designer's work. However, to our knowledge, there is no previous data on how to determine when a designer is willing to discuss their contributions to a piece of work or how to sense from their usage behavior that they are in a flow that should not be interrupted.

Figure 7: Explanation request pop-up pointing at the newest image in the mood board and asking why it was included. The image is significantly lighter than the previous images.

For this reason, I decided to implement a judgment based on the features of images rather than user behavior.

One possible problem with explanation requests is that designers might feel criticized by them, especially if the timing is based on a judgment of the images. If the AI asked for every image, the question would perhaps appear less critiquing as always asking communicates that it is not based on their behavior. Hence, it cannot criticize this behavior. On the other hand, this also communicates that the AI is unable to make its own judgment, bringing its competence and agency in question. Furthermore, asking questions without making own observations and judgment is likely to be seen as annoying. For this reason, I implemented a heuristic for judging when to make an explanation request.

The intuition of the heuristic is to make an explanation request when the newest images seems surprising based on the context. A surprise for the AI system could be seen as an image that it would consider a switch of strategy. It suggests images from the same or neighboring feature spaces. Hence, images that differ more than the scope of one strategy agents to the current context could be considered surprising to the AI system.

I implemented the function that decides if to make an explanation request using thresholds. These thresholds are based on the size of the strategy agents in the bandit system. These are 60°for hue and 0.5 for saturation and lightness. For contrast, I increase the threshold to 60°. Testing this methods showed that more explanation requests than intended were triggered by differing hues. To make the number more reasonable, I added a consideration of the contrast, so that the AI makes an explanation request for a different hue only of the mood board is otherwise harmonious. The intuition behind this is that a contrasting new color is not surprising on a mood board that is already colorful.

```
if makeExplanationRequest():
    open pop-up
function Boolean makeExplanationRequest():
    # big color distance in harmonic mood board
    if contrast(primary_color_new_image, primary_color_mood_board) > 60
            and mood board_contrast < 60:
        return true

    # unexpected saturation change
    else if |saturation_new_image - saturation_mood_board| > 0.5:
        return true

    # unexpected lightness change
    else if |lightness_new_image - lightness_mood_board| > 0.5:
       return true

    # unexpected change in contrast
    else if |contrast_new_image - contrast_mood_board| > 60:
```

```
        return true

    return false
```

The approach of asking when the images are differing significantly in their visual appearance might introduce a bias in the answers towards contrast. However, harmony might still be selected because the image can still have similar values for the other features. Content also remains a likely answer because the heuristic is not directly related to content. Even if a bias was introduced, the positive effect of reflection on the reasons for adding the image remains.

For the case that despite the approach of selective timing for making an explanation request, it appears at an inappropriate moment, the pop-up is easy to dismiss. It disappears by itself if the designer continues working. A click anywhere on the mood board is sufficient to close the pop-up. Furthermore, there will always be at most one explanation request visible to avoid cluttering the interface with multiple pop-ups. The system achieves that by first closing previous explanation requests before opening a new one.

4.3.3 Options and System Reaction

To explain the reason for adding the image, the designer can click one of the buttons: *content, harmony* or *contrast*. If the user explanation is *content*, the AI adjusts its association list to the content of the image. It queries a new list of associations with the search phrase that the designer used to find the image and replaces the previous associations with this list. This approach allows for a change of the topic to align stronger with the content of the explained image. If the designer selects *harmony*, the AI system sets the cost of choosing a different strategy agent in the bandit system to a higher value. A higher value for this cost makes exploitation less likely, and hence more suggestions will be harmonizing with the current mood board. Selecting *contrast* decreases the cost of selecting different strategies, with the effect that more suggestions have features differing from the current mood board. The designer can also choose not to react to the explanation request and dismiss the pop-up by clicking in the background.

5 User Study Methods

The evaluation of reciprocal explanation was conducted in the context of the whole May AI tool in collaboration with Janin Koch, the main creator of May AI. I assisted her in designing and setting up the study. The analysis of the data was conducted separately for May AI and this thesis. The description of the study setup in this chapter will focus on the information that is needed to understand the findings on reciprocal explanations. A more thorough description of the study can be read in the publication on May AI [30].

According to the previously defined research question, the goal is to investigate if reciprocal explanations can facilitate the co-creative work of a designer with an AI system. Thus, the aim of the analysis in the light on this thesis is to investigate if the anticipated benefits of the system and user explanations and its combination could be achieved in a design ideation tool.

Specifically, I investigate whether the explanations

- help to form a theory of mind for oneself and the AI (H1)
- inspire the designers to explore further properties of the images and the mood board (H2)
- lead to more alignment of the strategies of the AI and designer (H3)
- convey the ability and independent agenda of the AI (H4)

I further explore which role the explanations play for the designers during the interaction with May AI in our implementation. In case that the goals of reciprocal explanations can be reached partially, it is interesting to see how designers will react to this implementation, what already works in it, and to what extend improvements in the implementation promise reaching more benefits of reciprocal explanation. The study has a formative goal, exploring what makes a valuable explanation to a designer in ideation for future implementations. The design choices for reciprocal explanations in May AI were made based on the technical realization of the AI system and the limited available knowledge about explanations in design processes from previous work. The designers' assessment of the explanations and analyzing how they talk about their mood boards may help close this knowledge gap.

5.1 Participants

We recruited 16 participants from the target group, which are designers from different design disciplines. Some of the participants worked in several design disciplines. In total, the participants covered fashion, textile, industrial textile, graphic, industrial, material, furniture, interaction, urban, digital, service, strategic, product, and web design as well as architecture and fine arts. The designers' ages ranged from 26 to 39 with a median age of 33.5 years. Only designers with a minimum working experience of two years were recruited, which is a level of experience shown to be necessary to develop the critical level of abstract thinking which is a requirement for creating

effective mood boards [35]. These experienced designers are best able to judge the usability of the tool for this task. Less experienced designers cannot be expected to have the necessary knowledge of what makes an effective mood board and mood board creation process. Our sample of participants had two to 13 years of working experience in design with a mean of 5.79 (median 5) years. All participants had worked with the mood board method before. The participants volunteered under informed consent and were compensated with a ticket to a cinema.

5.2 Setup and Data Collection

The study was set up in a lab on a desktop computer, where the designers used May AI to create two mood boards. An examiner was present in the room to answer questions, take notes, and conduct semi-structured interviews. From the time of first log-in to the system to the end of the study, the screen was recorded with audio, capturing everything the participants did in the interface as well as their responses in the interview. The study took place in Finland and was conducted in English. The participants had diverse nationalities and had a good command of English as a first or second language.

5.3 Task and Procedure

At the start, the participants watched a video introducing the functionalities of the tool. The task then was to create two mood boards using May AI once with the AI and once without. One page long design briefs were given to the participants defining fictional but realistic customer tasks. The tasks were to develop a mood board for a new brand or sub-brand image to a well-known customer, one a bank, the other a grocery store. The participants had 15 minutes per task to complete the mood board. The briefs and conditions were counterbalanced to counteract learning effects for the tool and biases of the briefs on the process during either of the conditions. The work on each mood board was followed by a questionnaire (which was analyzed only for the May AI publication), a presentation of the resulting mood board, and a semi-structured interview. The procedure took one hour per participant.

5.4 Presentation and Semi-Structured Interviews

The participants had two minutes to present each of their mood boards. After the presentations, semi-structured interviews were conducted. The interview included questions about the satisfaction with their mood board and the process of creating it, the tool as a whole, and if it impacted their design process and how. The interview following the condition with AI support also included questions about the interactive behavior of the AI system and which impact the participants seemed to have on the AI and the AI on them. There were also questions particularly targeted at the effect of reciprocal explanations. The interview concluded with questions about the applicability of the tool in the participants' practice. The interview questions can be found in the appendices.

5.5 Data Analysis

I viewed the audio and video data of the mood board design process with the AI system, the presentations, and the interviews with each participant. During the mood board design process, I recorded how often each participant received an explanation request and how they reacted to it. Notes were also taken about the interaction with the suggestions and about comments made by the participant during the use. Transcripts of the mood board presentations were made for analysis of how participants explained why they included certain images or groups of images. If explanations occur during presentations from real designers, these may inform the content for future AI explanations. From the interviews, every mention of reciprocal explanations and their effects was recorded as a note. Furthermore, notes were taken about mentions related to our hypotheses such as reflection, understanding, trust, inspiration, alignment of goals, and agency. After a pass through all videos, the recorded notes were analyzed for reoccurring themes.

6 Results

6.1 Mood board Presentations

The participants presented their choices mainly talking about associations between the mood board topic or the target group and the motives or colors or other visuals of images. They talked about associations such as: “I picked this picture because I want to present the idea of nature and nice food” (P2), “I really was hesitating bringing this one here (pointing at an image with knitted fabrics) but bringing something soft and more unexpected to a blank visual image [...] to talk of different values and perhaps the easiness.” (P4), “I thought it needs a bit more color to bring some sportswear association” (P4), or “[...] the picture at the top with a person clearly in a busyness meeting in a suit. So combining the city life that they have with their business and that they have ideas.” (P5)

6.2 System Explanation

Most participants did not notice the system explanations and some read them only a few times. Only six participants said that they read the explanations by the system. One of them (P13) did not recall the content of the system explanations, leaving it questionable if they can be counted to the participants who read it. From the remaining participants, one (P9) did not know what the explanation meant. This participant recalled reading explanations that referred to small color contrasts (“this harmonious image matches your harmonious mood board”). They did not know what “harmonious” meant. The remaining four readers of the explanations also did not use them much. Participant one generally paid attention to the AI system only twice. Participant five turned to the explanations to find out how the AI came up with the suggestions and noticed that the explanations mostly referred to colors: “It suggested a lot about the color [...] when I added this picture there – the quite yellow one - it showed briefly a picture of these yellow swirls and said something about ‘this could fit the yellow picture you have’” (P5). Participant seven read only a few explanations because they thought the explanations were mostly the same. The six suggestions that they got were explained via associations, lightness, and saturation. Table 6.2 shows which participants read the system explanations.

6.3 Designers’ Understanding of the System

Most of the participants were uncertain about how the AI system worked. Nevertheless, many participants shared their assumptions about how the AI system arrived at the suggestions during the interviews. In these assumptions, I could identify a range of different speculations on which features the AI observed or which images or behaviors of the user the AI system took into account. Table 6.2 summarizes participants’ understanding of how the AI made suggestions.

Seven out of 16 participants mentioned colors or graphical style as a feature that the AI system used. For instance, participant two said: “At the very first time

Table 2: The rows show for each participant whether they read the system explanations, which features they mentioned as potentially relevant for the AI suggestions and which of their actions or work they mentioned that the AI potentially paid attention to as context for the suggestions.

Participant	Read system explanations	AI features	AI attention
1	yes	colors	background
2	no	colors	search terms, images on mood board, focused images
3	no	contents	images on mood board
4	no	themes, colors, graphical styles	images on mood board
5	yes	themes, colors	images on mood board
6	no	-	images on mood board
7	yes	colors, graphical styles	mood board
8	no	-	images on mood board
9	yes [3]	key words	search terms
10	yes	color associations, themes	images on mood board, search
11	no	-	-
12	no	contents	-
13	yes[4]	contents	recent images on mood board
14	no	colors	search, selected from AI
15	no	-	-
16	no	key words	images on mood board

it was very random and then the system starts to follow the colors in accordance to the things I pick" (P2). Participant 14 mentioned: "Yes the colors [...] I think after a while it started suggesting similar pastel tones and if you look these are all similar" (P14). Eight out of 16 participants mentioned that the AI system might have paid attention to key words, the content of the images, or overarching themes in the contents. For instance, participant three noticed that the AI system suggested many images with humans when they had chosen many images with humans for their mood board and participant 16 speculated: "Maybe for each picture it tagged me a certain amount of key words." (P16) Participant nine expressed their confusion about the connection between the keywords that they searched and the suggestions from the AI system: "I think it [their behavior] impacted [the system] but to me, it was not clear why it was affecting how it was and why keywords that I searched ended up pictures of certain images or patterns." (P9) Participant nine later also mentioned that they understood the connection of the content in suggestions at some point but complained that the AI system did not seem to take into account all of the keywords. Apart from participant nine, two more participants mentioned that they had "no idea" (P12) or were "wondering" (P8) where the suggestions came from. Three participants mentioned colors as well as themes as possible features in the AI such as participant four who considered a prioritized approach that to suggest according to a theme if possible: "Maybe if there is nothing theme-wise, that it can identify, it suggests based on color." (P4) Four participants did not talk about possible features. The four participants who read and remembered the content of the system explanations all mentioned colors as a feature that the AI paid attention to.

There were also various assumptions about which actions or artifacts the AI system observed as a context for the suggestions. The assumptions include roughly two categories; (1) the AI system paid attention to the mood board and (2) the AI paid attention to what happened in the search interface on the left and the AI interface on the right. There were some variations within both categories. Speculations included that the AI system observed the mood board as a whole (P8) or its background colors (P1), the most recent (P13) or all of the images on it (P2 - P6, P10, P16) or an image on it which they put into focus by clicking on it (P2). Clicking on images was how the tool allowed to select images to manipulate with the tools. A selected image was highlighted with a shade until it was deselected again. Within the category that considered the interfaces around the mood board, two participants mentioned their search terms (P2, P14), and two the search more generally (P10, P14). One of these also assumed the AI considered which of its suggestions they accepted (P14).

6.4 User Explanation

The amount of explanation requests that the participants received varied. Two of the participants (P4, P8) did not receive any. The maximum was seven explanation requests in one session (P15). For one participant (P10), the recording of the screen was interrupted during the process, so that no exact count exists. However, the participant remembered that the system asked them something and the participant could provide their opinion on the feature. The other participants all together received

Table 3: This table shows the number of explanation requests that each participant got and how they reacted to them.

Participant	Number of explanation requests	Reaction
1	4	Content: 2, ignored: 2
2	1	Content: 1
3	5	Content: 1, harmony: 3, ignored:1
4	0	
5	4	Content: 3, ignored: 1
6	4	Content: 2, contrast 1, ignored: 1
7	3	Content: 2, ignored 2
8	0	
9	5	Content: 3, harmony: 2
10	>1	Number and reaction unknown due to interrupted recordings
11	2	Content: 2
12	3	Content: 3
13	5	Content: 4, ignored: 1
14	2	Content: 1, harmony: 1
15	7	Ignored: 7. Did not notice them.
16	6	Content: 3, harmony: 2, ignored: 1

Table 4: This table shows whether participants (a) considered the user explanation a way to *guide the AI*, (b) saw a benefit through *reflection* or getting a *reminder* through the explanation request, (c) felt *interrupted*, (d) felt *critiqued* or (e) thought the user explanation had *non of these impacts*. Participants 2, 4, 8, and 15 are excluded because they did not see explanation requests.

Participant	Guide AI (a)	Reflection / reminder (b)	Interruption (c)	Critique (d)	no impact (e)
1	x				
3				x	
5		x	x		
6		x			
7					x
9	x	x			
10	x		x		
11					x
12		x	x		
13	x	x			
14		x			
16		x			

51 explanation requests (average 3.4). The most common answer was "content", which was selected 27 times. "Harmony" was the response to six explanation requests and only once "contrast" was selected. Seventeen explanation requests remained unanswered either because the participants dismissed, ignored, or missed them. Out of these, seven were ignored by the same participant, who, during the interview, said they did not notice them. In some cases participants accidentally dismissed the requests by clicking at the mood board before being able to react to them which, in some cases, led to confusion or the intention to retrieve them. For instance participant 9 accidentally dismissed their first explanation request and recalled that they were "confused what did I miss and did I do something wrong." In the recording it is visible that this participant clicked in the area where the pop-up disappeared. There were smaller issues with answering to the explanations requests. Such as participants wondering if it was possible to answer with several options (P13, P14). Furthermore, participant 14 wondered whether it was necessary to answer.

In general, there were divided perceptions and opinions on user explanations. A majority of participants (7 out of the 12 who remembered getting explanation requests) said that the explanations can have a positive influence by causing reflection or serving as a reminder for alternative strategies. There were also three mentions of feeling interrupted, and one felt criticized. Four participants mentioned that they thought that user explanations are a way to guide the AI system. Two participants did not see any positive or negative impact.

Regarding the impact of the reverse explanations on them, one commonly mentioned theme was reflection (P14, P6), that the questions provoke thinking about why they chose images (P13, P5) or about their strategy (P9, P16). “I think they had a positive impact because they give you a chance to reflect on why you make things” (P14), “I understood much better why I actually chose that picture.” (P6). Another common theme was the explanation request as guidance and reminder. There were comments such as “It listed some themes. So it could work as a reminder in a way.” (P12) or that the answer options indicate that they need to balance between harmony and contrast (P16). Participant 12 also saw a chance for teaching novices in mood board design that they need to think about different styles.

Some participants wanted to initiate user explanations to guide the AI system towards their strategy. In the recordings, it was observable at least by participants 9, 13, and 16 who right or double-clicked images and commented that they tried to trigger the pop-up. In the interview some participants also mentioned that they would like to “steer the AI to better suggestions” (P13), “train the AI in a way that would have made sense to me” (P10) or be “able to mark the relevance of other images with other features” (P13). There was also a mention that teaching the AI motivated the participant to answer the reverse explanation “maybe if I give more suggestions and answers then it would actually learn why I choose the pictures and what are my reasons” (P9). Another hint that teaching the AI motivated providing user explanation were comments such as that of participant one, who said that they “did not click it anymore” (P1) because they were “not expecting that it would give me anything” (P1) after they had answered content and the next suggestion of the AI was not aligned with the content of the explained image.

Some participants felt annoyed (P5) or interrupted (P10, P12) by the explanation requests. Participant 5 mentioned that the annoyance was due to hurry but they thought that the user explanations nevertheless helped them “It was important for me to define why I like the pictures so that was helping me in that matter. But at first, when the pop-ups came up I was like ‘Oh, I am in a hurry. I need to be ready’”. Also for participants 10 and 12, it was the flow that they felt got interrupted. Participant 12 found that interruptions should take place only if they got stuck while participant 10 thought that the purpose for the pop-ups was to train the AI system, which they were not willing to do actively during the workflow, but perhaps on their own initiative.

7 Discussion

Only six participants read the system explanations and only four remembered the content. The reasons for the rare usage are unclear. One obvious reason could be that the system explanations were hidden due to their visual design and placement in the user interface. Other possible reasons could be generally low attention paid by some participants to the AI interface due to a tight schedule for completing the mood board. Due to this, participants might have focused their attention only on actionable and faster to process parts of the AI system such as the suggested image and the buttons below it instead of the sentence of explanation. If this is the case, it is likely that given more time, more designers would notice the explanations. One participant said that they read the system explanations only a few times because they did not use the suggestions of the AI system (P1). This might have been the reason for overseeing the system explanations for a few more participants who also did not use the suggestions. However, the majority of participants did interact with the AI system, making other factors for overseeing the system explanations more likely for them. One of the participants also mentioned that they read the explanations only a few times because they thought they were always the same (P7). Hinting that the system explanations might have more value and get more attention if they were more versatile in phrasing and content.

With the little attention paid to the system explanations, limited amounts of data could be gathered on them and the combination of user and system explanations. However, all but three of the participants provided user explanations, resulting in a good amount of data on the more novel part of reciprocal explanations. The discussion in the following subsections on research question two on the additional value provided by combining two directions of explanations will be more speculative than on research question one about the individual directions, especially user explanations. For answering research question two, a new study with a more visible appearance and refined content of system explanations and a longer time to explore the user interface needs to be conducted.

7.1 Theory of Mind

Hypothesis one that reciprocal explanations help improve the theory of mind, in the sense that designers better understand the reasoning of the AI system and themselves through reflection, could be partially approved.

There is strong evidence that it aided reflection and at least some evidence that it helped understand the AI system. Despite the still high confusion among designers about how the suggestions worked the explanations probably could make it better for those participants who read it. The participants who remembered the explanations were among the ones that understood the AI system better. However, more data is needed to become more confident about this outcome.

7.1.1 Understanding of the AI System

There are many factors along the process of how the AI system selects suggestions that belong to understanding how the AI system reasons. First, there are the visual features that it uses to describe images and search terms that it uses to query them. Second, there is the context that it observes and feeds into the bandit system as a basis for the selection. Not understanding only one of them can already cause frustration. E.g. participant two first thought that the AI system used the image they most recently clicked at as context. In reality, it was always the whole mood board. This participant was confused because they had an incorrect mental model of the context. An example of not understanding which features mattered could be found with participant eight, who noted that the AI system suggested images after they had collected images on the mood board, hence they understood the context correctly. Nevertheless, they said that they were wondering about the relations to the suggestions.

Only four of the participants remembered the system explanations. All of these four understood that the AI system considered the mood board and colors. From the interview with participant five, it can be concluded the explanations helped learn that colors were a relevant feature for the suggestions. On the contrary, only three out of the remaining twelve participants who did not remember the system explanations identified color as a relevant feature for the AI. These differences between participants who remembered the explanations and those who did not, provide some evidence that the system explanations supported the understanding of how the AI system worked.

The results on participants' theories about how the AI system came up with suggestions show disagreement between participants on which features and which contexts were considered. Some of them are closer to how the system worked in reality, others were further off. It is not surprising that some participants who did not understand how the system worked explicitly mentioned that they found it confusing, that they had no idea where the suggestions came from or were at least uncertain. These results emphasize the need for more transparency so that users can build a more correct understanding of how the AI works and do not feel confused or frustrated about the AI system reacting differently to how they expected.

Some participants mentioned explicitly that they would use the system in their practice only if they knew the principles behind it. That is, not only how it comes up with the suggestions but also more generally where the images come from, who are the creators of the images, or which are the usage rights. This is an additional domain-specific aspect of transparency that goes beyond explaining how the algorithm works. When designing such a system for ideation for use in practice, this should be considered. These findings are also well aligned with previous findings by Koch et al. [29] who found that designers pay attention to these properties of materials when they browse the web to find inspiration.

Even though there is still a need for improvements towards communicating how the AI system works, there is some evidence that the system explanations helped the participants understand the AI system. As mentioned above, spending more time

using the system would allow users to pay more attention to the explanations and also get explanations using more different features. In the long run, this might lead to an increase in the transparency effect.

7.1.2 Reflection

The comments of participants who answered explanation requests show that user explanations helped them to think about the reasons for including certain images and to find out why they liked them. Thus, the hypothesis that explanations increase the understanding of one's reasoning through reflection could be approved.

7.2 Content for Explanations

This subsection discusses hypothesis two: Reciprocal explanation aids further-reaching exploration by reminding the designer to think about different features and strategies.

There is no evidence that system explanations supported the participants in thinking of different features or strategies. This is different for the user explanations. For them, there is evidence that, during the study, the three options *content, harmony* or *contrast* served as a reminder for considering different strategy options for the mood board. One especially experienced designer also mentioned that these questions could be used to educate designers who have no or little experience with mood boards (P13).

For being able to be reminded of something, the explanations need to contain information that the user did not think about at that moment, bringing up a surprise, teaching something new, or bringing something back into focus that they did not keep in mind. Hence, it relies on the content. One of the reasons for the system explanation not to remind the participants of different features could be that the visual features contained in most explanations are too obvious to further inspire an experienced designer. Despite this, they might still be important for the decision whether or not to include an image. Participant nine mentioned this: " I think I was focused on the content. But I think it is also always unconsciously focusing on the colors. It is very difficult to say why you choose pictures and what is your own practice of making picture choices." (P9) According to this, it still makes sense that the suggestions match visually to the mood board, similarly, visual features can be subject to the system explanations to reassure to designers that the visuals were considered. If visual features matter to designers, even if only subconsciously, this conveys that the suggestions are not random and that the AI system is able to consider these features. Knowing this is relevant for trust and understanding the AI (see above). However, expectations should be lowered on the inspirational impact of such features.

What could be more inspiring as the content of explanations are overarching themes, contents of images, associations of images, or color associations. The explanations from designers during the presentations of their mood board contained these kinds of abstractions. However, for reasons of trust, it might make sense

to reveal that the AI system also pays attention to visuals. As found above, this facilitates understanding the AI system and reduces unpleasantly surprising behavior that corrupts the perceived integrity of the AI system.

Mentioning these aspects during their mood board presentations might be a hint that these themes and associations were less obvious to the designers than the visuals. Furthermore, content was most often answered in reverse explanation. One reason for this might be that content was of higher importance for the selection. If the content is more relevant for the designers' decisions, it might also be more relevant for them to read about contents in the system explanations. Furthermore, there were mentions that designers were concentrated on finding good keywords (P3), found it helpful to get only a few images per search, so that they needed to think of more words (P13), emphasized that they found a specific word (P15) or started interacting with the AI suggestions when they did not have words (P5). This finding confirms previous research, which also found that finding words is challenging during ideation [54] This makes explanations containing themes, key words, or associations likely more valuable reminders than those containing colors.

7.3 Alignment of Goals and Strategies

This section discusses hypothesis three that reciprocal explanation aids the alignment of goals and strategies. No evidence backing this hypothesis could be identified in the interviews or recordings.

Some designers commented that they were influenced by the AI suggestions. This happened most likely through seeing the suggestion itself. The role of the system explanations is not known. No comments hinting to the alignment of strategies were given by any of the participants who read it. Possibly a different design and content of system explanations could affect the goals and strategies of designers. However, this needs to be tested.

The AI system tries to align its values every time a user provides an explanation. Value alignment can have positive and negative effects. Too little value alignment leads to contributions that cannot be integrated with the mood board because of lack of relation while too much value alignment leads to redundancy. It is hard to judge if too much or too little value alignment took place. There are hints in both directions as well as cases where it seemed to be at a good level. Examples are participant 13 who complained that the AI system repeated too much what they already chose, while participant one stated that the images were too unrelated and participant 16 was satisfied with the suggestions. This could be due to personal preferences that the AI system might can learn over time. In this case, the alignment would improve after longer usage. However, the need for more or less aligned strategies may also change over time as ideation alternates between divergent and convergent phases. Only one participant noticed that their answers had an impact on the strategy of the AI system, while one other said that they did not observe an impact. The rest of the participants did not state a clear position on this.

From the side of participants steering the AI system towards their strategy could be identified as a motivation to answer explanation requests. In that sense, a potential

for alignment of strategies in user explanations could be validated. However, with this study, I could not determine if the attempts by the AI system to align better according to the user explanations were an improvement for the ideation process. A study with a differently designed user explanation feature providing feedback about the effect of the user explanations on the AI system could deliver this information.

7.4 Communication of Agency and an Own Agenda

There is some but little evidence for hypothesis four that reciprocal explanation can communicate the agency and existence of an own agenda of the AI to the designer.

Many participants agreed that the AI system had its own agenda. However, mostly the comments about this referred to the AI panel as a whole, so that it cannot be distinguished how much of this impression was influenced by the explanations and how much by the suggestions and other parts of the AI panel. Some hints towards an increase of perceived agency through reciprocal explanations could be that there were participants who felt critiqued by the AI. Critique requires an own distinct opinion or agenda. An additional sign for agency could be that the questions asked by the system influenced the designers by making them reflect. Furthermore, comments were hinting that the pop-ups were surprising or unpredictable, a nature of action that also increases the scope of agency.

7.5 Design Improvements

Based on the findings, I would like to discuss possible design improvements or design choices to consider redesigning the reciprocal explanation feature. The improvement suggestions are discussed separately for each direction of explanation. Besides, consistency between the system explanations and user explanations should be preserved to ensure integrity and reciprocity. For instance, they should follow meaningful and similar principles, which underlines the existence of an agenda that conveys that the AI system only requests explanations that it would also provide itself.

7.5.1 System Explanation

The system explanations were overlooked by the majority of the participants. To some degree, it could help to change its visual presentation. It might, for instance, look more relevant if it was spatially closer to the suggestions or pointing at it similar to the explanation requests. This could visually communicate the reference to the images. Previously, it might have appeared like an arbitrary heading.

There was a problem for one of the designers to understand what the explanation was telling them by describing an image as “harmonious”. This was the translation for a low contrast i.e. small difference of the most dominant hues. To improve this kind of confusion, the natural language translations of contrast values should be reviewed. However, a good description of small differences in hues is challenging. For high contrast, the translation was “color contrasting”. Perhaps an analogous translation

such as “harmoniously colored” would be slightly clearer. Other possible opposites of colorful or contrasting, such as colorless, overlap with descriptions of medium lightness and low saturation. An alternative could be to provide explanations from a combination of these values. Focusing rather on which combination of features matches the term than translating from one feature value to a term. The color descriptions from the search term optimization could be adapted for this.

For making the explanations easier to comprehend, I made the design decision to limit the number of features explained at once. However, there are ethical reasons that favor more complete over more comprehensible explanations; it can be questionable to hide other features that played a role just for the sake of keeping it simple and convince to accept a suggestion. Making the whole reasoning explicit would be a more ethical solution because it provides a more integral basis for the decision whether or not to trust, rely on, and accept suggestions. My approach to completeness was to display varying information over time. However, a redesign could attempt to communicate more relevant features in a shorter time. This could be done by ensuring that the features alternate more quickly or by explaining using several features combined. Combining features could be rapidly realizable utilizing the functions for search term optimization.

However, with this approach, the comparison between the mood board and image suggestions might be more difficult i.e. whether suggestions match or complement the mood board could not anymore be determined by comparing just one feature. Furthermore, including features in the explanations which are not visible, such as the hue of barely saturated color. These factors make explanations with combined features non-trivial. Alternatively, visual representations of features might also explain the relationship between the mood board and suggestions in a more understandable way e.g. juxtaposing dominant colors. That would have the additional benefit that the color analysis leading to the feature vectors would be more explicit in the explanation. The disadvantage of this approach is the lack of verbal communication, which could aid a designer’s reflection on verbal descriptions for their mood board.

The slicing of the design space into strategy agents in the bandit system underlying May AI differs from the feature selection thresholds and feature translation borders. The technical reasoning behind the coarse slicing into uniform and regions is to enable exploration. The alternative would have been dynamic slicing which could easily lead to hard to distinguish small slices [30]. I decided to set custom thresholds and borders between translations based on subjective perception. The HSL color space is in accordance with hardware such as screens and printers but has drawbacks in the uniformity of hues and the brightness of colors for human perception [26]. This made different partitions necessary. However, aligning explanations and the underlying algorithm increases transparency. This should be considered. More accurate alignment might be possible e.g. becoming even more fine-granular in the translations so that additional borders could be introduced in the same positions as in the bandit system. Alternatively, options for changing the algorithm could be explored to match the wording of the user group or a more perceptually uniform color space could be used. For instance, the CIE Lab color space is designed to fit closer to perceived color differences [57]. Its color wheel is used to support the creation

of color swatches in modern tools for graphic design such as Adobe Illustrator [5]. However, this color space is rarely supported in libraries for color analysis and hence difficult to implement.

A shift to more associations and content related explanations might make sense in order to yield a higher value of the explanations as a source for inspiration. As mentioned above visual features might be too obvious for an experienced designer to continue being inspirational. It might still make sense to communicate them to show that the AI system considers these nonetheless important features. However, expectations should be lowered on the inspirational impact of this information. Additionally, a stronger focus on key words or associations would make sense. Other than evaluating how well an image matches visually, finding words for their design is a challenge during design ideation [54], making key-words a promising candidate for inspiring explanations.

There seems to be a conflict between explaining with obvious visual features for improved transparency and explaining with inspirational key words. To get both effects, an explanation strategy change over time could be introduced. Many participants said that they did not see the system explanations. It is unclear if they did not see them because they did not seem relevant or if they were not salient enough visually. In both cases spending more time with the interface could make more users notice them. Seemingly less relevant and less visible content would be looked at eventually. Increasing the time even more, the users would see explanations using several features, hence could improve their understanding more of how the AI works. This could be continued until the users learned how the system works. At this point the purpose of explanations shifts from educating to providing inspiring information. At this point, the content of the explanations should change e.g. to associated key-words or color associations.

7.5.2 User Explanation

There were complaints about the irregularity of the explanation requests because they appeared for some pictures but not for all (P13). The timing of the reverse explanations is currently based on the visual features of the image and the mood board. If they differ by a certain threshold, it is interpreted as "surprising" to the AI system with the consequence that it asks for the reasoning behind including the image. It is purposefully balanced to not interrupt for every image. However, based on the comments from the participants, the following alternatives for the timing could be meaningful as well.

If it was possible to determine when a designer is in a state of "flow", it could be avoided to request explanations in these moments and instead ask for an explanation when they are "stuck". However, it is unknown how to judge when that is the case based on observable behavior. More research would be needed to detect the feeling of flow or being stuck from the interaction with a graphical user interface. Instead of visual features of an image, also the content could be used to determine how surprising an image is to the AI system. For instance, the AI system could find associations

[5] https://helpx.adobe.com/illustrator/using/color.html, retrieved 26. April, 2020

with the search term used to find the image and compare these associations with its current association list. If there is no overlap, it could be considered surprising. Such an approach would align well with a general shift towards utilizing key-words more in the reciprocal explanations. Another approach could be to leave the timing up to the user. Several participants tried to open the pop-up and more participants than expected were motivated to steer the AI system through user explanations. This motivation should be supported by allowing user initiative for their explanations.

The results also suggest reconsidering the answer options for user explanations. The AI system asked based on large differences in the features. It might not be a surprise that harmony was chosen infrequently as a response due to that. However, harmony was selected much more often than contrast. Contrast was selected only once, which might be a sign that contrast is not usually a goal of designers when they create a mood board. This might also indicate that harmony and contrast mean something else to designers than visual similarity or difference. Especially harmony is a somewhat ambiguous term. For instance, apart from similar in color, it can also refer to a pleasing combination or balance. Even adding a contrasting color can lead to an overall more pleasing whole [39]. The ambiguity is not necessarily negative as leaving room for reflection based on an own interpretation might have value as well. However, the interpretation needs to be transparent to the AI system to enable meaningful reactions to the answers. The answer options should be more clearly specified or the adjustment of the AI strategy needs to be adjusted to the various possible interpretations of the answer option.

The options for user explanations were limited to just three: content, harmony, and contrast. However, more versatile options for answers might be feasible and helpful for reflection and steering the AI system. The current limit to three options was based on the assumption that answering should be as easy as possible. Increasing the number of options or even allow free text was assumed to interrupt the workflow for too long. However, the positive reaction of the participants to being able to steer the AI system encourages me to propose more versatile opportunities for user explanations. For instance, after selecting content they could be given the option to define the content by writing a word or view and edit the key-words that the AI system would add to the association list. Especially increasing the granularity for explaining via content could be meaningful given that 80% of the given user explanations were content and that some participants answered the explanations requests to guide the AI system in a certain direction regarding the content.

One more reason for providing the option to review key words before the AI system adds them to the association list is to provide feedback about the consequence on the answer. This would increase transparency about how the AI system handles key words, which was a point of confusion among designers. Allowing to revert their answer or altering the key-words before submitting the explanations to the AI system would further empower their communication with the system.

8 Limitations and Future Work

The user study was conducted with two conditions, one with the AI system and one without it. There was no control condition with the AI system but without reciprocal explanations. Hence, no definite conclusions about which effects were caused by the presence of the AI system as a whole and which were added by the explanations. This was due to limitations in resources. It was not feasible to invite 16 professional designers for an additional study. This was appropriate to gather the first qualitative data about the benefits and drawbacks of reciprocal explanations and inform improvements for the implementation. Future work could implement an improved version according to the results and control the impact of reciprocal explanations against no explanations and explanations only in one direction.

Another limitation of the study is its temporal restriction; each of the mood boards was created within 15 minutes and every designer created only one mood board with the help of the AI system. The short time per mood board resulted in an unnaturally rapid design process, which allowed for less time for exploration of design ideas and most likely also the user interface. With respect to the system explanations, this leaves the question open whether more participants would notice it over time and which effect it would have on the co-creative ideation process. The short time also might have affected the user explanations. Many of them were not answered, which might have been influenced by a lack of time. More time would also allow to reflect longer, which could potentially shift user explanations to more subconscious options.

The perception of user explanations might have been affected by the short time frame as well. Apart from the feeling of interruption, which was discussed above 7, participants could gather only limited experience with user explanations. The average number of explanation requests per participant was below four. This number could rarely provide participants the opportunity to experiment with the effects of their answers on the behavior of the AI system.

The study was conducted with professional designers with a minimum of two years of practical working experience and previous usage of the mood board method. This selection criterion was based on the observation that the abstract thinking skills develop with practice and reaching the necessary level for creating effective mood boards takes about two years /citelucero2012framing. These experienced designers could provide more knowledgable feedback about the applicability of the tool for design practice. However, as it was also pointed out by a participant who teaches design courses P(13), the tool and especially the explanation could be helpful for novices in mood board design. To evaluate this potential, future work could conduct a study with less experienced users, such as design students.

Explanations are just one part of creative discussions. According to Dorta et al. [11], collaborative ideation in design teams consists of iterations of (1) naming or constraining, (2) negotiation, (3) decision making and moving. Each iteration starts with naming a specific part of the design to be discussed and optionally pointing out constraints such as budget or time. This opens a phase of negotiation which entails making propositions, questioning, and explaining. The negotiation is concluded

with a decision on whether to agree or disagree with a proposal. Optionally the group performs a move action, i.e. conducting a change on the design. Reciprocal explanations allow for explanations and some degree of naming by referring to a contribution to be explained and questioning by issuing explanation requests. However, essential research questions on how to further increase the capacity of AI systems to engage in collaborative ideation are open for future research. They include how it can maintain a negotiation, incorporate previously mentioned constraints, propositions, and explanations in its own suggestions, explanations, and questions. Furthermore, future work could investigate to what degree and under which conditions an AI system should attain the capacity to make decisions or perform move actions.

The study results emphasize the strong focus of designers on key words and associations during the search of inspirational material as well as the challenge to find such words for their ideas. This finding is well-aligned with the insight that the designers continuously translate between visual and verbal representations of design ideas [54]. However, it seems like the search for visual material and verbal expressions take place simultaneously. Future work could explore possibilities of supporting the search for expressive verbalizations for a given design problem at the same time as exploring visual material.

Finally, reciprocal explanations were developed for co-creative tools for ideation, which can be regarded as a subcategory of mixed-initiative systems. However, it goes beyond the scope of this thesis to test the transferability of reciprocal explanations to other mixed-initiative systems and how they can benefit from a stronger reciprocal communication of reasons behind various agents' activities.

9 Conclusion

This thesis presents the novel technique of reciprocal explanations which aims to improve the communication of human designers and AI systems while both contribute to joint ideation work. In human groups, the exchange of explanations plays an important role in enabling collaboration. It aids reflection, mutual understanding, the alignment of goals, and the joint utilization of diverse ideas and perspectives. Previous explanation techniques have not enabled these advantages due to a lack of reciprocity. Mostly, AI systems only explain their own contributions and cannot receive explanations from users. If they do, these explanations mostly neglect the needs of the user by serving only the training of the AI system. Reciprocal explanations aim to allow a more even partnership by introducing explanations in both directions and emphasizing the link of each explanation to a contribution to the ideation process - which is of equal importance for both partners. System explanations provide the reasoning behind contributions from the system, intending to make the system transparent and providing additional inspirational material. User explanations provide the opportunity to the user to reflect on their reasoning and communicate their goals to the AI system by explaining their contributions.

Reciprocal explanations were integrated into an AI aided tool for mood board design, a popular method for ideation. The tool contributed by suggesting images that could fit the mood board. System explanations were realized in textual form linking one of the features present in the system's suggestions to the current mood board. Users could explain by answering to explanation requests which asked them why they included an image. The tool was evaluated with 16 professional designers who were experienced in mood board design. During interviews following the mood board creation with the tool, we gathered insights about their perception of the explanations.

It could be validated that user explanations aid reflection and can serve as a reminder to also consider different strategies. Furthermore, I found that participants' motivation for explaining often was to steer the AI system. This is an interesting finding suggesting that user explanations could be a valued tool for aligning strategies between users and AI systems. Unfortunately, only four of the participants read and understood the system explanations, leading to limits in the findings on the value of system explanations and the inclusion of two directions of explanations. However, the findings suggest that there is a need for increased transparency especially for those participants who failed to notice the explanations.

This motivates a future continuation of the work also on system explanations and a second study with an improved implementation and study design. Based on the observations of the study, this could include a redesign of system explanations towards more explanations with key words and a study that enables the users to spend more time exploring the interface of the tool. Furthermore, a future study could include a control condition without reciprocal explanations. This was not possible in the scope of this thesis due to the difficulty of recruiting expert designers for an extended amount of time. Furthermore, future work could investigate complementary techniques that

enable AI systems to engage in creative discussion and the transferability of reciprocal explanations to other types of mixed-initiative systems.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.

[2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

[3] Michael J Baker. Collaboration in collaborative learning. *Interaction Studies*, 16(3):451–473, 2015.

[4] Eric PS Baumer, Jordan Sinclair, and Bill Tomlinson. America is like metamucil: fostering critical and creative thinking about metaphor in political blogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1437–1446, 2010.

[5] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*, 2017.

[6] Siddhartha Chaudhuri and Vladlen Koltun. Data-driven suggestions for creativity support in 3d modeling. In *ACM SIGGRAPH Asia 2010 papers*, pages 1–10. 2010.

[7] Erin Cherry and Celine Latulipe. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(4):1–25, 2014.

[8] David D Chrislip and Carl E Larson. *Collaborative leadership: How citizens and civic leaders can make a difference*, volume 24. Jossey-Bass Inc Pub, 1994.

[9] N Davis, C Hsiao, Kunwar Yashraj Singh, Brenda Lin, and Brian Magerko. Quantifying collaboration with a co-creative drawing agent. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(4):1–25, 2017.

[10] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 196–207, 2016.

[11] Tomás Dorta, Yehuda Kalay, Annemarie Lesage, and Edgar Pérez. Design conversations in the interconnected his. *International Journal of Design Sciences and Technology*, 18(2):65–80, 2011.

[12] Tomás Dorta, Annemarie Lesage, Edgar Pérez, and JM Christian Bastien. Signs of collaborative ideation and the hybrid ideation space. In *Design Creativity 2010*, pages 199–206. Springer, 2011.

[13] John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37, 2018.

[14] Vegard Engen, J Brian Pickering, and Paul Walland. Machine agency in human-machine networks; impacts and trust implications. In *International Conference on Human-Computer Interaction*, pages 96–106. Springer, 2016.

[15] Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. Twenty years of creativity research in human-computer interaction: Current state and future directions. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 1235–1257. ACM, 2018.

[16] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[17] Milene Gonçalves, Carlos Cardoso, and Petra Badke-Schaub. Inspiration choices that matter: the selection of external stimuli during ideation. *Design Science*, 2, 2016.

[18] Barbara Gray. Collaborating: Finding common ground for multiparty problems. 1989.

[19] Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 1884. NIH Public Access, 2018.

[20] Marti A Hearst, J Allen, C Guinn, and Eric Horvitz. Mixed-initiative interaction: Trends and controversies. *IEEE Intelligent Systems*, 14(5):14–23, 1999.

[21] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016.

[22] Tom Hewett, Mary Czerwinski, Michael Terry, Jay Nunamaker, Linda Candy, Bill Kules, and Elisabeth Sylvan. Creativity support tool evaluation methods and metrics. *Creativity Support Tools*, pages 10–24, 2005.

[23] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166, 1999.

[24] Giulio Jacucci, Anna Spagnolli, Jonathan Freeman, and Luciano Gamberini. Symbiotic interaction: a critical definition and comparison to other human-computer paradigms. In *International Workshop on Symbiotic Interaction*, pages 3–20. Springer, 2015.

[25] Shu Jiang and Ronald C Arkin. Mixed-initiative human-robot interaction: definition, taxonomy, and survey. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 954–961. IEEE, 2015.

[26] George H Joblove and Donald Greenberg. Color spaces for computer graphics. In *Proceedings of the 5th annual conference on Computer graphics and interactive techniques*, pages 20–25, 1978.

[27] Ben Jonson. Design ideation: the conceptual sketch in the digital age. *Design studies*, 26(6):613–624, 2005.

[28] Andruid Kerne, Eunyee Koh, Steven M Smith, Andrew Webb, and Blake Dworaczyk. combinformation: Mixed-initiative composition of image and text surrogates promotes information discovery. *ACM Transactions on Information Systems (TOIS)*, 27(1):1–45, 2008.

[29] Janin Koch, Magda Laszlo, Andres Lucero Vera, Antti Oulasvirta, et al. Surfing for inspiration: digital inspirational material in design practice. In *Design Research Society International Conference: Catalyst*. Design Research Society, 2018.

[30] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. May ai?: Design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 633. ACM, 2019.

[31] Janin Koch and Antti Oulasvirta. Group cognition and collaborative ai. In *Human and Machine Learning*, pages 293–312. Springer, 2018.

[32] Samantha Krening, Brent Harrison, Karen M Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55, 2016.

[33] Joseph CR Licklider. Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1):4–11, 1960.

[34] Henry Lieberman. User interface goals, ai opportunities. *AI Magazine*, 30(4):16–16, 2009.

[35] Andrés Lucero. Framing, aligning, paradoxing, abstracting, and directing: how design mood boards work. In *Proceedings of the designing interactive systems conference*, pages 438–447. ACM, 2012.

[36] Lucia Mason. Collaborative reasoning on self-generated analogies: conceptual growth in understanding scientific phenomena. *Educational Research and Evaluation*, 2(4):309–350, 1996.

[37] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Organizational trust: A reader*, pages 82–108, 2006.

[38] Donald Michie, David J Spiegelhalter, CC Taylor, et al. Machine learning. *Neural and Statistical Classification*, 13(1994):1–298, 1994.

[39] Jill L Morton. *Color logic*. Colorcom, 1998.

[40] Nils Johan Nilsson. *Artificial intelligence: a new synthesis*. Morgan Kaufmann, 1998.

[41] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[42] Hugo Gonçalo Oliveira, Tiago Mendes, and Ana Boavida. Co-poetryme: a co-creative interface for the composition of poetry. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 70–71, 2017.

[43] Florian Pinel and Lav R Varshney. Computational creativity for culinary recipes. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 439–442. 2014.

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[45] Mark O Riedl, Jonathan P Rowe, and David K Elson. Toward intelligent support of authoring machinima media content: story and visualization. In *Proceedings of the 2nd international conference on INtelligent TEchnologies for interactive enterTAINment*, page 4. ICST (Institute for Computer Sciences, Social-Informatics and . . . , 2008.

[46] Sebastian Robert, Sebastian Büttner, Carsten Röcker, and Andreas Holzinger. Reasoning under uncertainty: Towards collaborative interactive machine learning. In *Machine learning for health informatics*, pages 357–376. Springer, 2016.

[47] Thomas R Roth-Berghofer and Jörg Cassens. Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In *International Conference on Case-Based Reasoning*, pages 451–464. Springer, 2005.

[48] Ben Samuel, Michael Mateas, and Noah Wardrip-Fruin. The design of writing buddy: a mixed-initiative approach towards computational story collaboration. In *International Conference on Interactive Digital Storytelling*, pages 388–396. Springer, 2016.

[49] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[50] Frode Sørmo and Jörg Cassens. Explanation goals in case-based reasoning. In *Proceedings of the ECCBR*, pages 165–174, 2004.

[51] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143, 2005.

[52] Shashank Srivastava, Igor Labutov, and Tom Mitchell. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1527–1536, 2017.

[53] Gerry Stahl. Group cognition, 2006.

[54] Anne Tomes, Caroline Oates, and Peter Armstrong. Talking design: negotiating the verbal–visual translation. *Design Studies*, 19(2):127–142, 1998.

[55] Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. Idea expander: Supporting group brainstorming with conversationally triggered visual thinking stimuli. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, page 103–106, New York, NY, USA, 2010. Association for Computing Machinery.

[56] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. Accelerating human-in-the-loop machine learning: challenges and opportunities. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2018.

[57] Xuemei Zhang, Brian A Wandell, et al. A spatial extension of cielab for digital color image reproduction. In *SID international symposium digest of technical papers*, volume 27, pages 731–734. Citeseer, 1996.

A Interview Questions - No AI

A.1 Outcome Quality: Scale 1-7

How satisfied were you with the moodboards?
How useful did you find it to explain your case?
How novel did you find it?

If this interview followed the second moodboard:
If 1 is the first moodboard and 7 the second – where is your preference?

A.2 Tool in general:

How did you experience the functionalities of the Tool?
Did it hinder you to do anything and if yes what was it?

A.3 Interaction:

Can you tell me how you experienced the interaction with the Tool?
Did you feel that the Tool impacted your Moodboard process?
How? / What hindered you?
Did you feel that your behavior impacted the system?

A.3.1 Agency:

How would you characterize the system within the making process?
If you have to describe it in a visual image – how would you describe the role of the AI?

A.4 Suggestion:

Did anything of the system created an "aha – moment"?
Did you think that the system pointed you in different directions than you intended?

A.5 Applicability

Would you use such a system in your work practice and when?
Y: What for in specific; N: What needs to be changed?

B Interview Questions - With AI

B.1 Outcome Quality: Scale 1-7

How satisfied were you with the moodboards?
How useful did you find it to explain your case?

How novel did you find it?

If this interview followed the second moodboard:
If 1 is the first moodboard and 7 the second – where is your preference?

B.2 Tool in general:

How did you experience the functionalities of the Tool?
Did it hinder you to do anything and if yes what was it?

B.3 Interaction:

Can you tell me how you experienced the interaction with the Tool?
Did you feel that the Tool impacted your Moodboard process?
How? / What hindered you?

B.4 AI Interaction

Can you reflect on how your behavior impacted the system suggestions?
How much did you feel the system understood your aim?
Did you have the feeling that the system had its own agenda to follow?

B.4.1 AI agency:

How would you characterize the AI within the making process?
If you have to describe it in a visual image – how would you describe the role of the AI?

B.5 AI Suggestion:

Did you have the feeling that the AI suggestions were meaningful?
Did you experience an "aha – moment" while using it?
Did you think that the system pointed you in different directions than you intended?

B.5.1 AI reflection

Where you asked something by the system about your choices?
What impact had the questions the system asked you on your behavior?
How useful where the explanations of the system?

B.6 Applicability

Would you use such a system in your work practice and when?
Y: What for in specific; N: What needs to be changed?
Where do you see the potential of such systems and where the limitation?