ARNY: AN INTERACTION MODEL BASED ON EMOTIONAL FEEDBACK FOR AN AI-BASED CO-CREATIVE DESIGN SYSTEM

by

Sarah Abdellahi

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2021

Approved by:

_____
Dr. Mary Lou Maher

_____
Dr. John Gero

_____
Dr. Heather Lipford

_____
Dr. Aidong Lu

PREFACE

The data collection for the second evaluation study of this dissertation started in late Feb 2020 and was paused shortly after in March 2020 due to Covid 19 pandemic that forced closure of our research labs. The original data collection plan called for recruiting 50 demographically homogeneous participants, dividing them into two groups of 25, and assigning them to individual study sessions with two study tasks in counterbalanced order. A statistical analysis was planned to compare frequency of each affect category between treatment and control tasks in order to detect statistical significance. The researcher also planned to compare the participants' ratings of the tasks as well as the expert review scores for task outcomes to determine whether statistically significant differences existed. To analyze the qualitative feedback collected during the debriefs, the researcher was planning to perform a thematic analysis of retrospective reports for the 50 participants and compare the patterns identified for treatment vs. control conditions.

After the face-to-face data collection was stopped due to the pandemic, it was not possible to continue data collection in an online setting due to context and environmental factors having a huge impact on an individual's experience of affect. As a result, the researcher was unable to collect more data and continue the initial plan of thematic and statistical analyses. As the Covid-19 impact on face-to-face interactions continued and the researcher could not resume the data collection, she was forced to opt for narrative analysis of the data as an alternative.

It should be noted that the claims made as a result of this study findings are tentative and based on the limited data available. Data could be examined in more depth if higher data volume were available or the initial data collection had planned for a narrative analysis

that accounted for collecting more details about the participants' individual differences and included different follow up questions for understanding of their personalities and the impact of personality on interactions with the systems.

ABSTRACT

SARAH ABDELLAHI. Arny: An interaction model based on emotional feedback for an AI-based co-creative design system. (Under the direction of Dr. M. Maher)

Co-creative collaboration is a creative collaboration of two or more agents working together on a shared creative product. Effective co-creative collaboration is a combination of interactions and contributions of the task collaborators. In human-human collaboration, gestures, verbal communications, and emotional responses are among the general communication strategies that shape the interactions between the collaborators and enable negotiation of the contributions. Emotional feedback allows human collaborators to passively communicate their stance about the experience and convey their perception of the process without distracting the flow of the task. In human-human co-creative collaboration, participants interact and contribute to the task based on their perception of the collaboration over time. However, perceiving the cognitive state of the user to determine the dynamics of collaboration and decide what the agent should contribute to the artifact are two primary challenges of building effective co-creative Artificial Intelligence systems (Abdellahi et all, 2020). In response to these two challenges, the following thesis statement is presented:

Using knowledge of human emotion in human-AI co-creative collaboration can improve the user satisfaction of the collaboration experience and quality of the collaboration outcome.

This thesis focuses on establishing dynamics of co-creative collaboration between a human and a co-creative AI agent through an emotion-based interaction model for co-

creativity. After the introduction in chapter 1 and reviewing the related background in chapter 2, the 3rd chapter of this dissertation presents an interaction model for Arny V1, a co-creative system designed to explore the research questions of this thesis. Arny V1 interaction model was studied as part of an exploratory study with the Wizard of Oz setup that is discussed in chapter 4 of this document. The modified version of the model, Arny V2, was then deployed and analyzed to confirm the thesis statement, as well as identify possible improvements. Studies of Arny V1 and Arny V2 confirmed this dissertation's thesis statement that consideration of affect in human AI creative collaboration can improve the satisfaction of the experience as well as the quality of the outcome. This research also identified valence and engagement as two emotion dimensions beneficial for designing affective co-creative AI agents.

DEDICATION

To Baba, Maman and Sorour for giving me love since day one!

To Pishu, the best friend and writing companion!

ACKNOWLEDGEMENTS

My entire PhD journey would not have been possible without my advisor, Dr. Mary Lou Maher, who guided, encouraged, and mentored me diligently throughout.

I also wish to thank UNC Charlotte faculty Dr. John Gero, Dr. Heather Lipford, and Dr. Nicholas Davis for their input on this research and all their help during my PhD.

I am grateful to UNC Charlotte Staff, especially Sandra Krause for all her support and kindness.

I would like to thank my labmates particularly Safat Siddiqui, Jeba Rezwana, and Ali Almadan for their contributions to data collection of this research.

My special appreciation goes out to my family: my parents and my sister, for their support and encouragement as I wrote this dissertation and throughout my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: INTRODUCTION

Computational creativity is a multidisciplinary intersection of artificial intelligence, cognitive science, art and psychology that studies the application of computer technologies to replicate, study, stimulate and enhance human creativity (Jordanous, 2014; Rouse, 2019). One of the areas where computational creativity has directed a significant amount of attention over the past few years is in creative collaboration research. Creative collaboration is specifically valued as it leads to the emergence of ideas far beyond what could be accomplished by individuals (Sawyer & Dezutter, 2009).

1.1 Designing Co-creative Systems

Numerous researchers are interested in designing systems referred to as co-creative systems that are capable of creative collaboration with --humans due to the benefits creative engagement presents in different areas from healthcare to education and industry (Sanders & Stappers, 2008; Kohler et al, 2011; Makhaeva et al, 2016).

To develop an effective co-creating system, two primary challenges need to be addressed: 1) determining the interaction dynamics of the collaboration, e.g., whether the system should lead, follow, wait, and 2) determining what the agent should contribute to the shared creative product and why. Human collaboration negotiates both factors through verbal and non-verbal social cues. Humans communicate their stances about these matters through conscious and unconscious body gestures, facial expressions and verbal feedback. Such feedback reflects the human feeling and the value judgment of each stage of the collaborative process and guides the interaction and contribution. However, current designs of co-creative systems lack such negotiation mechanisms. Designing a feedback

mechanism that effectively informs the co-creative agent how it should interact to increase creative engagement and fluid interaction is a challenge this thesis investigates (Abdellahi et al, 2020).

1.2 Emotion as Feedback Mechanism for Co-creative Agents

Emotions align with different human cognitive states and allow humans to reflect and communicate with their collaborative partners in order to calibrate the collaboration dynamics and their behavior. In a computational setting, emotions are an ideal candidate for feedback since they are passive, meaning that the user does not need to explicitly click or say anything. Unlike other negotiation methods such as voting buttons and verbal feedback, the passive characteristic of emotional feedback allows negotiation over the matters to happen without distracting the flow of the process. In contrast to methods such as verbal feedback, passive emotional feedback does not require the user to learn or adjust to any new method of communicating feedback (Abdellahi et al, 2020). This dissertation states that:

> *Using knowledge of human emotion in human-AI co-creative collaboration can improve the user satisfaction of the collaboration experience and quality of the collaboration outcome.*

While there is a large amount of existing work on affect based interaction, there is relatively little research about interaction models in the field of co-creativity, and especially affect based interaction in the existing co-creative systems (Abdellahi et al, Glines et al, 2021). Therefore, the outset for designing emotion based Human-AI co-creative interaction is to determine the design parameters of an emotion-based interaction model for co-creativity as suggested by the first research question of this dissertation:

RQ1: What are the design parameters of an emotion-based interaction model for co-creativity?

With this research question, this dissertation draws on a design for an emotion-based co-creative AI agent. This design is then used to check the thesis statement by answering the research questions below:

RQ2: What is the human perception of co-creativity with emotion detection?

RQ3: How are outputs different between interactions with and without emotion detection?

Considering the sparse literature on the topic and the complex nature of co-creative interaction this research follows an exploratory approach to iteratively design and revisit an emotion-based co-creative AI agent. This dissertation refers to this co-creative system as Arny, and uses it to pursue RQ2 and RQ3

1.3 Arny

The idea of considering emotional feedback in co-creative AIs that led to designing Arny is rooted in the researcher's observations during a previous study of human-to-human collaboration. In that study, each participant was paired up with a facilitator from the research team to collaborate on a set of drawing tasks. As a follow-up to the drawing tasks, the participants had to reflect on their collaboration experience during a retrospective process (Abdellahi et al, 2019). Thematic analysis of that case study exhibited that different individuals' responses to similar drawing contribution types that were explored by the facilitator varied in many ways. Despite these differences all participants reported a collaboration strategy structured around reflecting on the facilitator actions outside the

drawing canvas through verbal feedback and body language. This behavior pattern is in line with what was previously suggested by the literature about human interactions during the collaboration and how it impacts the collaborator's contribution (Sawyer, 2014). The same study found that individuals who were more focused on interacting and communicating their perceptions and feedback expressed higher levels of satisfaction with the quality of the collaboration.

These observations inspired this research to consider a Human-AI collaboration model that considers not only the collaborators' contribution to the task, but also a method of interaction to communicate perceptions and expectations similar to what is followed in Human-Human collaboration (Abdellahi et al, 2020). In addressing the need for an interaction channel between the user and the AI system, this dissertation presents the use of affect and facial expressions to facilitate communication and modulate the interaction dynamics and behavior of co-creative AI agents.

Since the design and deployment of such co-creative AI will be dependent upon the response to RQ1, four more specific sub-questions were identified to guide the model design:

1. Emotion stimuli: What stimuli can trigger emotions during a collaboration?
2. Capturing user's affect: How to capture and represent the target emotions in users in relation to the emotion stimuli?
3. Affect interpretation: After those stimuli are identified, which dimensions of the induced emotions by them can result in a meaningful interpretation of the user feedback and expectations?

4. Interaction Model: How to respond to the affects triggered by each stimulus in the course of collaboration?

This dissertation research begins by constructing Arny V1, the first iteration of an emotion aware co-creative system designed based on 1) conjecture 2) experiences from previous human-human co-creative research 3) literature on affect based interaction, and co-creativity.

This dissertation will review the related literature in Chapter 2, and then introduce Arny V1 in Chapter 3. RQ1 and RQ2 are explored in Chapter 4 through a study shaped around users' interaction with Arny V1. Chapter 5 introduces Arny V2, an improved design for Arny inspired by Arny V1 study findings. Chapter 6 discusses a case study around Arny V2 study findings and responds to RQ1,2&3. Chapter 7 concludes the dissertation by discussing limitations and opportunities for future research.

Chapter 2: BACKGROUND

Co-creation is usually referred to as the creative collaboration of two or more agents working together to produce a shared creative product. However, when the common definition of co-creation has a strong focus on the collaborations being around the product, a few definitions go further than that and look at co-creation in relation to agents' personal emotions and perceptions. For example, Miyake (2002) refers to co-creation as "co-emergence of real-time coordination between two or more agents sharing subjective space between different persons".

Defining co-creation with respect to sharing the subjective space emphasizes the importance of social and interactional aspects of co-creation in addition to the resulting artifact. The required interactional considerations while designing co-creative systems is discussed further in this dissertation. To distinguish co-creative systems from other types of computationally creative systems, it is first necessary to define the term.

2.1 Computationally Creative Systems

Over the past few years, creative collaboration has been introduced as a potential solution for many needs and contexts. To bridge the gap between users' needs and designers' mental models (Sanders & Stappers, 2008), some companies are using virtual collaboration environments like Tele-Board (Gumienny et al, 2013) and Second Life avatar-based virtual world (Kohler et al, 2011) which allow design contributions from their user community and remote employees. Co-creative activities such as Handlungsspielraum have been designed for educational purposes (Makhaeva et al, 2016). Moreover,

therapeutic applications are developed around co-creative practices to support a number of therapeutic purposes, including those related to autistic patients, depression patients, and aging in place (Paulus et al, 2010, Makhaeva et al, 2016). An example of such work is the RHYME project: A set of “co-creative tangibles” designed to improve the health and quality of life of people with severe disabilities through motivating play, communication, and co-creation. The introduction of the applications for creative collaboration raises the need for a wide range of co-creation systems capable of supporting creativity and timeless availability, which reduce social barriers for individuals struggling with direct social interactions (Morgan et al, 2014).

To distinguish co-creative systems from other forms of computationally creative systems, Davis et al (2015) introduced a model to categorize existing computational creative systems into three different categories based on the approach used by each category for supporting creativity (Figure 1): “Creativity Support Tools”, “Generative Systems”, and “Computer Colleagues”.

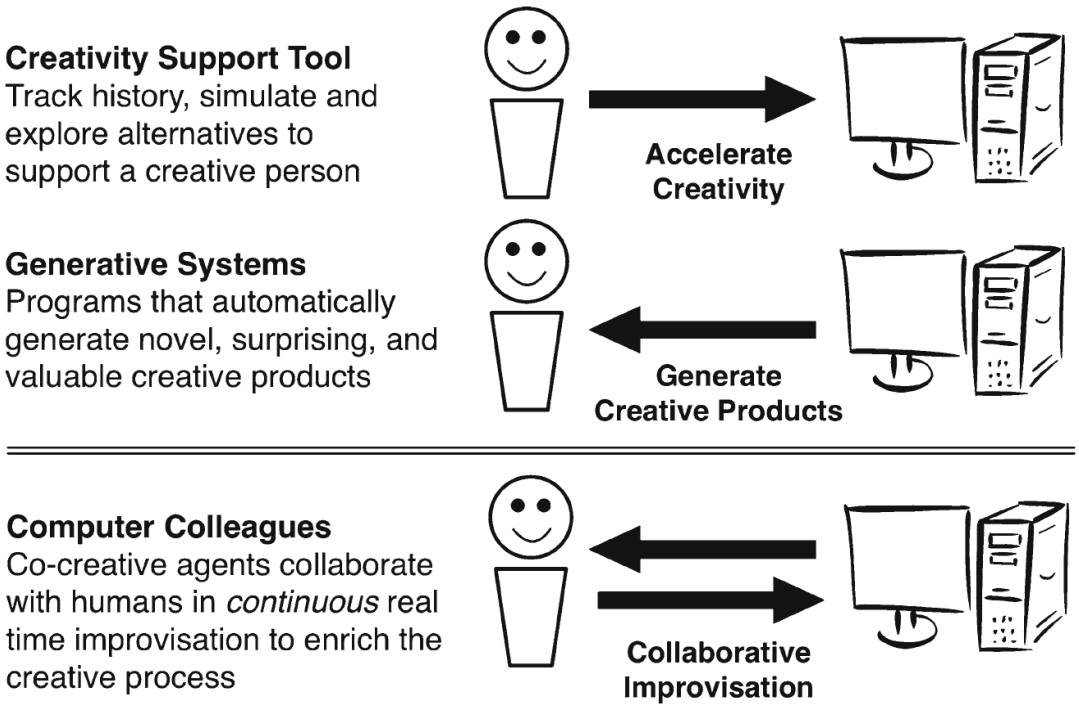


Figure 1- Computationally Creative Systems categories (Davis et al, 2015)

Creativity support tools

Creativity support tools are a subset of the field of computational creativity that improve the user's' ability in creative tasks, improve the results they get from a given set of abilities, support learning about a domain, or allow the user to experience aspects of the creative task that they would not be able to experience otherwise. Adobe creative toolset is a known example of creativity support tools (Davis et al, 2015; Shneiderman, B. 2007).

Generative systems

Unlike Creativity Support Tools, the focus of Generative Systems is not on providing the user with a new experience of the creative process. Instead, these systems are AI systems capable of autonomous generation of creative products. Deep Dream

Drawing Generator is an online example of Generative systems (DiPaola & McCaig, 2016; Davis et al, 2010).

Computer Colleagues

Computer Colleagues are the most complicated model of computationally creative systems as they focus on real-time improvisational creative interactions between the user and system with the goal of co-creating a creative product. These systems are a hybrid of Creativity Support tools and Generative systems as they focus on working alongside humans as a partner or colleague. The ultimate objective for Computer Colleagues is reaching a state comparable with human-human creative collaboration. The Computer Colleague model, referred to as the Co-Creative System in this proposal, is the subject of this research. ViewPoint interactive Dance system and Drawing Apprentice collaborative drawing system are two examples of Computer Colleagues (Jacob et al, 2013; Davis et al, 2015).

Among the three models, the Computer Colleagues is of interest to this thesis since its constant interactivity with the collaborating users makes it capable to support both the benefits from collaboration and creativity. Throughout this thesis, the Computer Colleagues model is referred to as the Co-creative system. Parts of this content has been published in Abdellahi et al, 2020 in International Conference on Human-Computer Interaction and Abdellahi et al ,2020 in ICCC.

2.2 Affect for Co-creation Interactions

Due to the critical role of emotion in human interactions, the field of Human-Computer Interaction strives to make human-computer interactions more spontaneous and human-like by including emotions in the process. In human-human interactions "Emotional expressions are crucial to development and regulation of interpersonal relationships" (Ekman, 1999). Research has shown individuals with facial paralysis experience great levels of difficulty developing and maintaining even casual relationships as they are incapable of expressing emotions effectively. This observation about human interactions inspired the current trend of considering affect in systems design (Ekman, 1999).

Currently, in different contexts of system design, the responsibility for adaptation has shifted from the user to the system itself. In other words, it is not expected that the user understands the system and adjusts to the interactions' complications. Instead, it is the system that has to understand the users' and provide a low barrier method of interaction. Based on this philosophy of human-computer interaction, system design takes into account users' characteristics and reactions (Hudlicka, 2003). The philosophy has been adopted to the level that Hudlica 2003 states: "In some communities we no longer even speak of users and machines as separate entities, but rather of collaborative systems...The synergy of technological and methodological progress on one hand, and changing user expectations on the other, are contributing to redefining of the requirements for what constitutes affective and desirable in HCI".

When it comes to collaborative systems however, consideration of affect is much more than a user's expectation and desire. In a collaborative system when the user and the

system contribute to a shared task, consideration of affect has a significant influence on the interaction between the participants and consequently the main sensemaking flows of the collaboration. Emotions elicited by stimuli events trigger responses in the participants and allow them to adapt to the collaboration (Scherer, 2005; Sawyer, 2014).

Kellas and Trees (2005) refer to two distinct sense-making processes in open-ended improvisational interaction, such as collaborative drawing or having a conversation: 1) functional sense-making that determines the content generated for a particular turn, e.g. choosing to draw a house or pattern, or choosing which words to say, and 2) interactional sense-making that structures and maintains how the interaction is unfolding through time i.e. the interaction dynamics, such as turn taking, turn length, and the overall rhythm of interaction. Participatory sense-making occurs when there is a mutual co-regulation of these two sense-making processes between multiple participants, i.e. both participants are adapting their responses to each other and working to maintain an engaging interaction dynamic that supports the mutual exchange (De Jaegher & Di Paolo, 2007) (See Figure 2).

When participatory sense-making occurs, the interaction can take unpredictable paths and new ideas can emerge by traversing through new conceptual spaces and generating responses to unpredictable queries. In human collaboration, collaborators naturally co-regulate their sense-making processes through awareness of their collaborator's judgment of their contribution at each point of time. This awareness allows the collaboration path to intuitively shape itself. Awareness of a collaborator's emotions during a collaboration allows the participants to validate their actions from their collaborator’s point of view and use this awareness to proceed with the participatory sensemaking (Eligo et al, 2012). The meaning structures built by the interactional sense-

making process guide the interaction forward by suggesting what can be added next given the history of the interaction. Also, interaction patterns are developed that circumscribe the type and amount of content to be generated at a given time. For example, when getting to know each other, people often employ a question-and-answer interaction pattern that suggests when each person should ask another question to keep the interaction moving. The same concept of having a pattern can be true for collaborative drawing –interaction dynamic patterns emerge such as call and response, mimicry, mutual building, antagonism, and transformation.

Once an interaction pattern is established through awareness of affects and context, cognitive resources can be turned from interactional sense-making to functional sense-making, and the participant can focus solely on generating a response to their partner in line with the latest interaction pattern being employed rather than generating a new contribution from scratch. This process can repeat and the observed changes in the affect or shared product during the collaboration could direct the participant to choose to re-engage in interactional sensemaking to come up with a new way of interacting and establish a new interaction pattern.

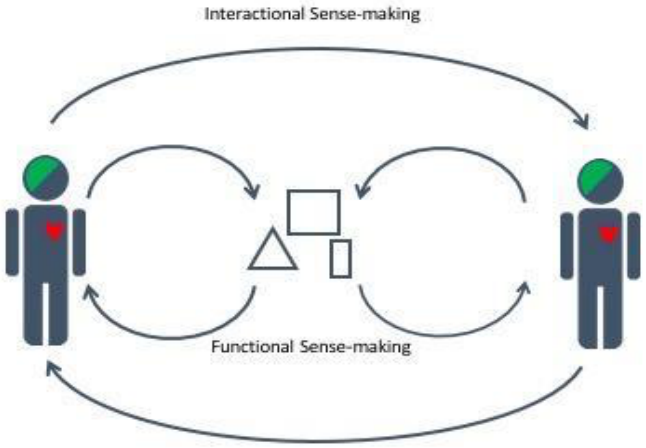


Figure 2- Sensemaking processes in an open-ended improvisational interaction

2.3 Characterizing and Measuring Emotions

To incorporate affect in the design of computing systems, it is required to 1) characterize and 2) measure human affect. Several methods of characterizing and measuring affect have been introduced and implemented in affective systems research and designs. Psychological studies introduce different definitions of emotion and suggest a variety of models on how affect could be characterized (Scherer, 2005; Ekman, 1999; Kleinginna and Kleinginna, 1991; Russell, 1980). However, all models for characterizing affect can be described as two main groups: Categorical models of affect, where emotion/affect is described in relation to a list of primary affects referred to as basic emotions, (Ekman,1999), and Dimensional models of affect, where emotion/affect is characterized in multidimensional spaces based on different measurable aspects such as intensity and positiveness. Circumplex model of affect, where affect is characterized on a two-dimensional space of pleasure-displeasure axis referred to as valence and an Arousal-calm axis, is one of the most commonly used examples of such dimensional models (Russell, 1980). The Geneva emotion wheel, on the contrary, is one of the famous categorical models of emotion characterization that tries to describe different emotional experiences in relation to a primary set of 20 basic emotions. In an attempt to describe emotions other than the main basic emotions, the Genova model provides a mapping of its basic emotions on a two-dimensional space of control and pleasantness. The basic emotions in the Genova model are of the extreme values possible for each category and when category sub-emotions are lighter control/pleasantness versions of these main emotions. The Genova emotional wheel and many other categorical models of emotion are not widely used as they are ineffective at describing emotions beyond their basic set of emotions

(Scherer, 2005). Another challenge with Categorical models of emotions is the impact of cultural factors on emotional responses to events. Recent studies have shown that what is considered as basic emotions is defined based on the western profile of emotion expressions and does not match main emotional responses experienced and expressed in other cultures such as Asian culture (Jack et al, 2012). The dimensional models of categorizing emotions, in contrast, have been commonly used due to their capacity of describing emotional change over time in terms of simple measurable values. This characteristic and all the existing emotion evaluation technologies based on different dimensions (Doerrfeld, 2015), make the dimensional model of characterizing emotion a suitable candidate for describing affect in the context of affective system design. Figure 3 shows a diagram of how emotion is being characterized on the Genova wheel of emotions and the Circumplex Dimensional model.

Emotion has been measured using a variety of methods including self-report affect evaluation, biometrics and bodily symptom measurements, facial and vocal expressions, and action tendencies (Appelhans & Luechen, 2006; Gross & Levenson, 1993). Among these methods, emotion measurement based on facial expressions is one of the most commonly used methods of emotion capturing and evaluation. Affect evaluation based on facial expressions is preferred by some researchers as its passive approach allows minimum distraction of the participants from the task. Darwin argued that facial expression method of affect evaluation could be used universally as most emotions are expressed in the same way on the human face regardless of cultural and racial backgrounds (Magdin and Prikler, 2018; Darwin and Prodger 1998). Moreover, this method of affect evaluation does not

require specific hardware components and could be developed over commonly used technologies such as personal computers and smartphones.

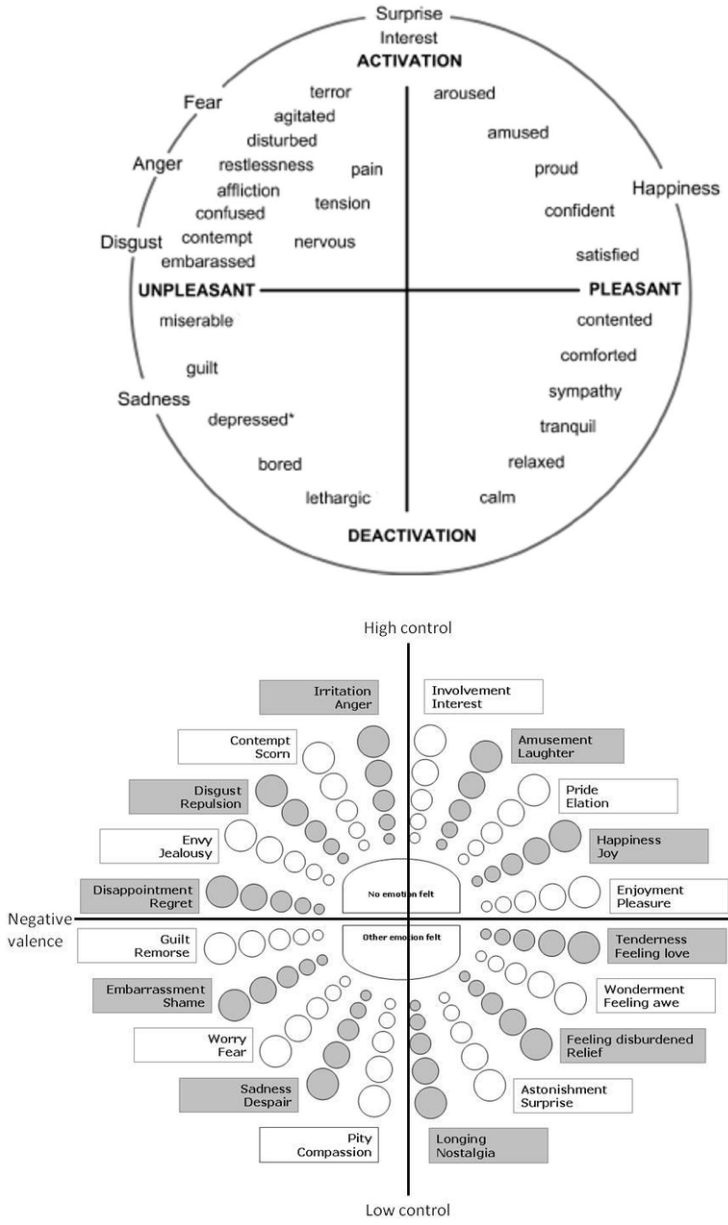


Figure 3- Sample presentation of emotions in the Circumplex model of emotion (top), Genova emotion wheel (bottom) (Plutchik,1991; Sacharian et al, 2012)

Chapter 3: ARNY V1, AI-BASED EMOTION REACTION INTERACTION

When it comes to designing an affective system, the primary decision to be made is the "role of affect" within the system. Human system interactions refer to affect in a variety of ways. In system design, affect can be taken into account by recognizing user affect, adjusting to the user's affective state, generating affective states within an agent, or a combination of these options (Hudlicka, 2003). The interpretation addressed by Arny, the emotion aware Co-creative drawing system that is designed for the purpose of this thesis, is recognizing the user's affect and adapting to the user's emotional state.

While most Co-creative AI systems only focus on the functional sense-making in order to collaborate with their human colleagues (Figure 4-left), Arny, follows a model similar to Figure 4-right to incorporate the partner's emotion in the system design and include the interactional sensemaking component to its sensemaking cycle.

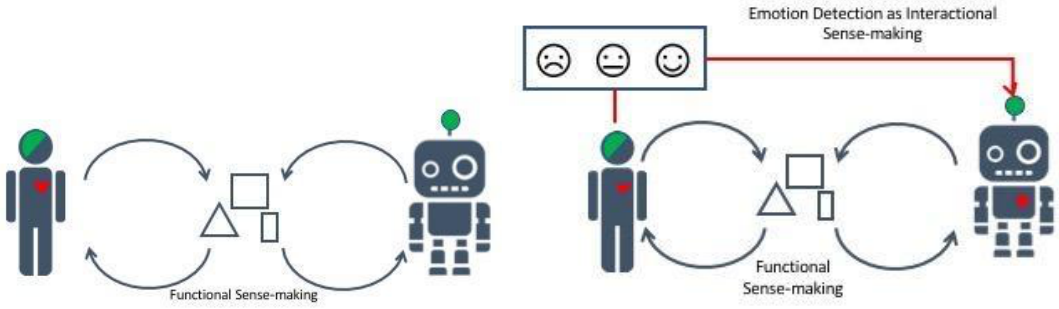


Figure 4- Human AI Co-creation without consideration of emotional feedback (left) Vs in the presence of emotional feedback (right)

The four sub-questions of RQ1 listed in Chapter 1 illustrate four major decisions for designing and evaluating the first iteration of Arny. This chapter describes Arny's approach to system design according to these four points: 1) Emotion Stimuli during co-

creation, 2) Capturing user emotions, 3) Emotion interpretation during co-creation, and 4) Collaboration strategy (model) based on the awareness of user emotions. Parts of this content has been published in Abdellahi et al, 2020 in International Conference on Human-Computer Interaction and Abdellahi et al ,2020 in ICCC.

3.1 Co-creation & Emotion Stimuli

Emotions are generally elicited by two types of stimuli: 1) External stimuli, when outside events trigger emotion, such as natural causes or behavior of other people, and 2) internal stimuli, when one's own behavior can be the event that triggers the emotion, such as with pride or shame (Scherer, 2005). Emotions experienced in a collaboration set up could be triggered by actions of the collaborator, collaboration environment, or by one's perception of their own contributions. As this research aims to understand the user's perception and emotions triggered by actions of the co-creative AI, Arny's co-creative design experience is structured to allow us to distinguish external stimuli triggered by Arny's actions and disregard affect triggered by other types of stimuli.

Arny utilizes a turn-taking pattern in the collaboration in order to uncover the emotional source. Using a turn-taking pattern, allows to distinguish emotions evoked by Arny in the user from intrinsic emotions evoked by how participants perceive their own contributions since the two are not contributing simultaneously.

Similar to collaborator's behavior, the complexity of the platform is an extrinsic event source that could lead to changes in participants' affective state. In order to differentiate the affect triggered by Arny from other possible sources, the drawing platform used as the context for Arny's collaboration is chosen to have low complexity and be easy

to learn. This drawing platform only provides the user with three simple tools, a pen tool, an eraser tool, and a select and resize tool. Due to its simplicity, the platform allows for a minimal level of mistakes or slips that could lead to frustration, confusion, or annoyance when interacting with the environment.

3.2 Capturing User Affect

Since the role of affect in the design of Arny is to investigate the user's emotional state and adapt to it, one fundamental aspect to investigate in Arny's design is affect recognition. User affect can be measured through a variety of biological and psychological methods including heart rate, facial expression, body gestures, diagnostic tasks and self-report (Picard & Daily, 2005; Kapoor & Picard, 2001). Each of the mentioned methods has advantages and disadvantages in terms of accuracy, ubiquitous and intuitiveness. The emotion recognition method suggested for Arny uses the facial expression method of emotion recognition as it provides a sufficient level of accuracy for the purpose of this research. Moreover, using this method, emotions can be measured in real time and in a passive manner, without distracting the user or interrupting the user-system interactions.

Arny uses the user's emotional state reported to it by Affectiva- a 3rd party facial expression emotion detection tool. The reported emotion by Affectiva is then interpreted by Arny, and considered in the model of interaction between Arny and the human collaborator.

Affectiva is a real-time facial expression recognition toolkit currently available as an add-on to the iMotion biometrics evaluation package. Affectiva captures the participant's facial expressions using a simple webcam. To code the captured expressions, Affectiva uses the "Facial Action Coding System (FACS)" which is the most widely used

model of coding facial behaviors [McDuff et al, 2016]. Affectiva can code different factors of facial expressions based on a large data set of facial expressions previously coded manually by FACS experts for training purposes (McDuff et al, 2016). Affectiva's automated facial coding system is comprised of the following components as presented in Figure 5:

1. Facial detection and extraction of Facial features
2. Classification of facial action points
3. Emotion expression modeling

Figure 5-Affectiva Facial Coding Model Components (McDuff et al, 2016)

1. *Facial Detection* and *Extraction of Facial features*: Facial detection in Affectiva happens through the Viola-Jones face detection algorithm which is outside the scope of this dissertation. After the face detection stage, the face is divided to 34 rectangle landmarks to be used in the later components of the model (Magdin and Prikler, 2018; McDuff et al, 2013; McDuff et al, 2016)
2. *Classification of Facial actions:* For facial action classification purposes, Histogram of Oriented Gradient (HOG) is extracted from the face regions defined

by landmarks. Then SVM classifiers trained by 10000 manually coded facial images are used to score the facial actions.

3. *Emotion Expression Modelling:* Anger, Disgust, Fear, Joy, Sadness, Surprise, and Contempt are the emotions recognized by Affectiva based on the combination of detected facial actions. These emotions are presented by a score from 1 (absent) to 100 (present) the Affectiva interface.

In addition to the emotion modeling, facial actions detected by Affectiva are used to present a valence value between 100 to +100. Table-1 presents a list of expressions used for the calculation of the positive and negative valence values by Affectiva.

Increase Positive Likelihood	Increase Negative Likelihood
Smile Cheek Raise	Inner Brow Raise Brow Furrow Nose Wrinkle Upper Lip Rais Lip Corner Depressor Chin Raise Lip Press Lip Suck

Table 1- The Valence metric likelihood based on facial expressions (Affective help center, 2019)

Affectiva also reports engagement value of the participant as a weighted sum of Brow raise, Brow furrow, Nose wrinkle, Lip corner depressor, Chin raise, Lip pucker, Lip press, Mouth open, Lip suck and Smile facial expressions (Affective help center, 2019). Figure 6 is a screenshot of valence, engagement, and joy expression reported by the Affectiva interface.

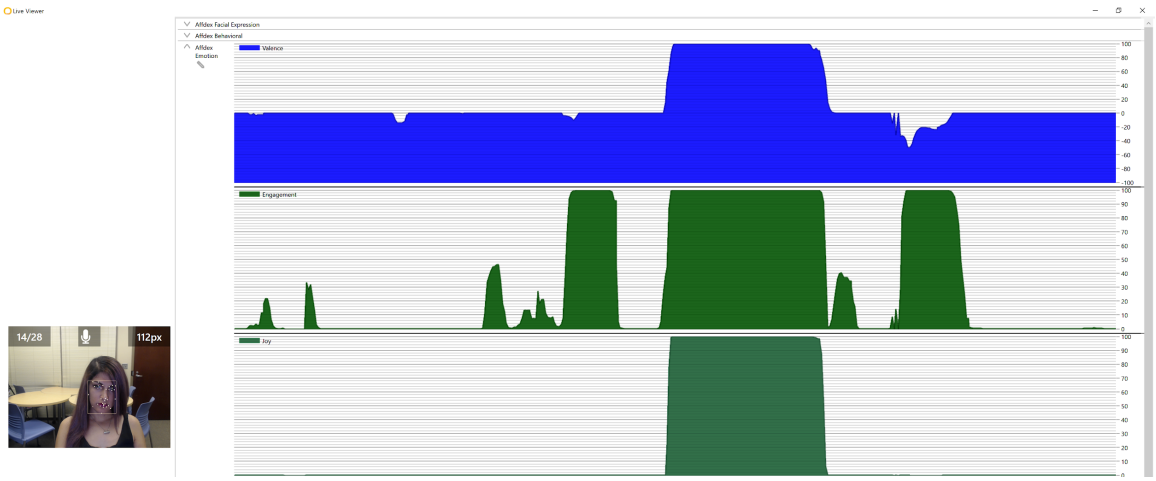


Figure 6- Valence, Engagement and Joy values reported by the Affectiva interface

3.3 Emotion interpretation during co-creation

As mentioned in section 3.1, Arny's design focuses on emotions that are triggered in the user through a specific collaborator's action. Arny correlates the captured affect with a specific AI action through its turn taking strategy and focuses on the user's judgment of one specific action at a time. The Diagram in Figure 7 shows emotions that can be evoked externally through collaborators' actions in a co-creation context. Emotions with positive valence such as amused, interested, and confident are emotions reflecting a positive value judgment of the collaborator that could motivate further engagement of the subject. However, emotions on the negative side of the continuum such as disconnected, bored, annoyed and dissatisfied represent the subject's negative value judgment of the collaboration, which can presumably result in loss of interest and decrease of engagement over time. Arny V1's design focuses on categorizing the users' value judgment of each contribution as positive, neutral, or negative without investigating the specific interpretation of each emotional response. In other words, Arny V1 only interprets the

user's emotional response as: did the user feel positive about an action, did the user feel negative about that action, or the user didn't have any significant feelings about the action.

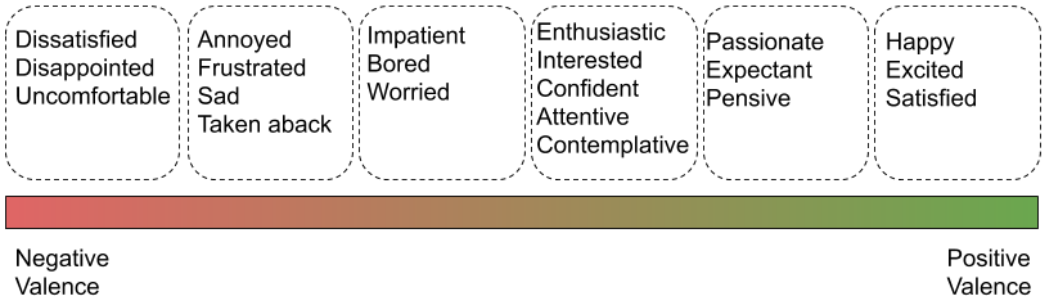


Figure 7- Valence correlation with affects elicited by collaborator

3.4 Interaction Model

In an emotion-based interaction, the decision on what is the preferred state to adopt in response to each identified emotional state depends on the specific context and system subjective (Hudlicka, 2003). As an example, a decrease in stress level is preferred in some situations while a rise in stress level is preferred in others. For instance, a training system where frustration results in the user becoming distracted is not like a flight automation system where you need to keep the user stimulated to prevent boredom and a lack of vigilance on the part of the pilot (Matthews et al, 2000). Previous researches however, have not reported much about preferences for an interaction between a human and a co-creative AI agent. The primary interaction model for Arny's design is constructed by using conjectures and observations of collaborative human action. A revised interaction model for co-creative collaboration is then proposed based on the preliminary evaluations of this model.

Arny's interaction model (Figure 8) includes two basic components, a collaboration model and a set of response rules. The collaboration model discusses the basic structure of

the collaboration, which allows us to capture the user's emotional response to each specific action performed by Arny as well as an overview of what is collected as affective feedback. The response rules on the other hand go deeper and discuss how Arny responds to each specific feedback type with consideration of the collaboration rules.

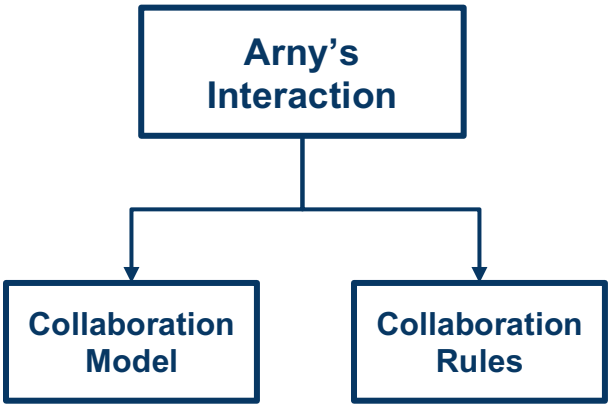


Figure 8- Arny's model of interaction components

3.4.1 Arny V1 Collaboration Model

The main objective of Arny during the collaboration is to maintain the positive affect in the user by selecting the next action in a way that won't provoke a negative impact on their emotions. The collaboration model is structured to allow Arny to track the source of different emotions expressed by the user and respond to them. As a way to track the source of emotions, the collaboration model follows a turn-taking pattern. With the turn-taking pattern, Arny can identify the user's emotional response to each specific contribution without confusing it with the participant's intrinsic emotional response. The turn-taking pattern suggested by Arny's collaboration model breaks down the collaboration process between the user and Arny into collaboration cycles. Each collaboration cycle includes a user turn followed by an Arny's turn. On each user turn, the user is allowed to contribute

one drawing object to the shared drawing space. In each cycle, Arny responds to each user's contribution according to the user's contribution in the beginning of the cycle and their emotional feedback in response to the Arny's contribution in the last cycle.

When collaborating with Arny, users are allowed to choose their contributed object freely and may converge to what was drawn in the previous turn of drawing or diverge from the previous drawing. A converging action in this definition refers to a contribution with the intention to follow the same mental model as the collaborator. Diverging actions, on the contrary, refer to the contribution of a new element that has the potential to cause a conflict between the AI and the user's mental models [Fuller & Magerko,2010]. In the context of collaborative drawing, a converging contribution can be a contribution with visual or conceptual similarity to the collaborator's actions, while cognitive divergence refers to a distance in terms of visual similarity, conceptual similarity or both. The present study only assessed conceptual similarity as a means for identifying convergent or divergent actions.

On each Arny turn, it contributes none or one drawing object to the shared drawing space. This drawing object could be converging to what was drawn in the previous turn of drawing or diverging from the previous drawing. For Arny, the decision on when to converge, diverge, or pass is made according to the collaboration rules and based on the captured user's emotion in the current cycle on the user's turn and the user drawing input from the previous cycle. Figure 9 demonstrates the collaboration model followed by Arny v1.

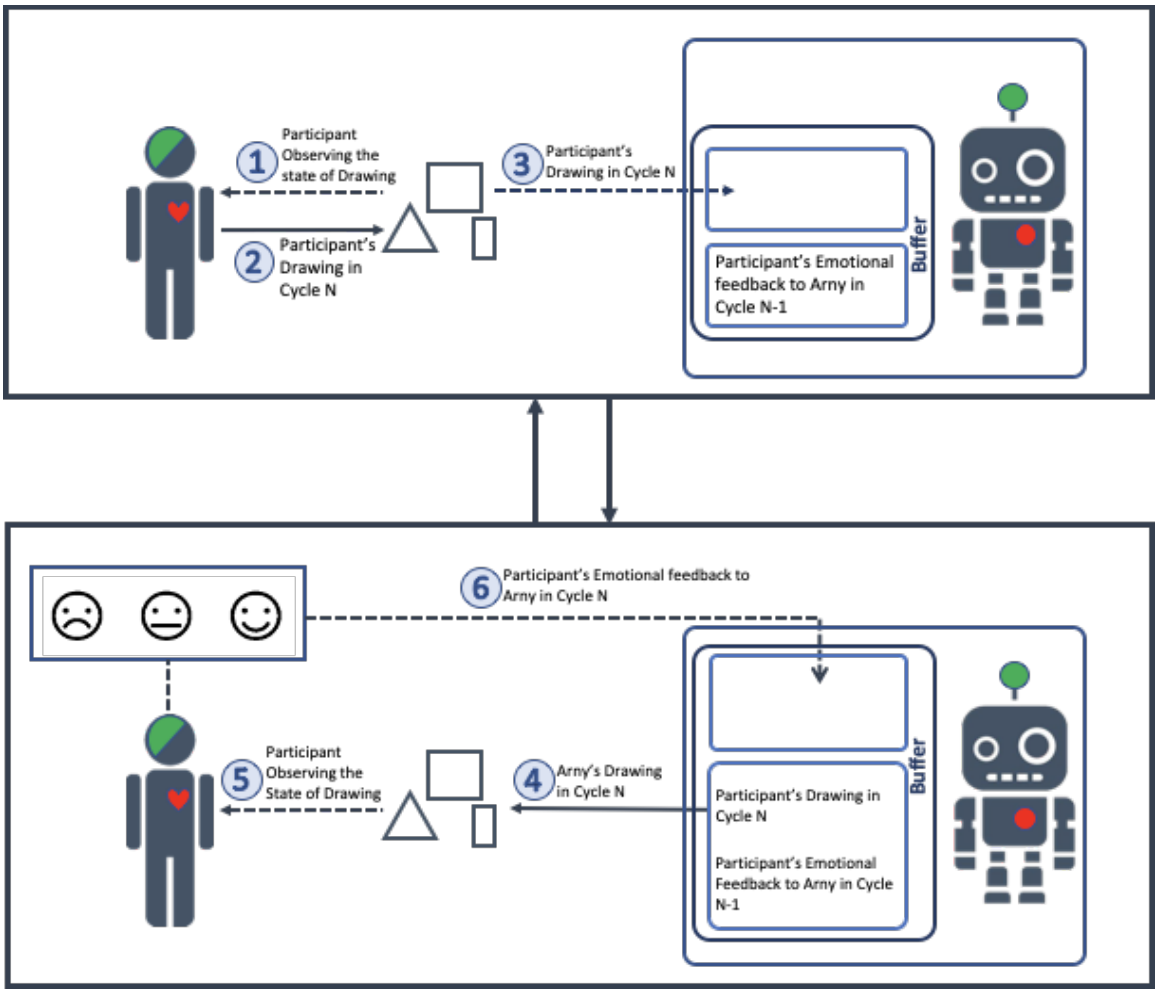


Figure 9- Arny's Collaboration Cycle- Participant Turn (top) and Arny Turn (bottom)

Figure 9 presents an Arny's collaboration cycle. A user turn is represented in the top box from Figure 9. This part of the cycle consists of three stages marked by numbers 1-3. The turn starts with stage 1 when the user observes the current drawing and decides on a contribution. At stage 2, the user makes a contribution to the shared drawing. Stage 3 happens after completion of the user's drawing, at this stage, the user 's drawing contribution is perceived by Arny and buffered in Arny's memory to be used in the next turn.

Arny's turn happens right after the user's turn in each cycle. Arny decides on when to converge, diverge, or pass based on the previously recorded emotional feedback of the user. Each Arny turn consists of three stages labeled from 4-6 in the lower box of Figure 9. Since the overall process is made by loops of the cycles, on each Arny turn, the user's drawing data from the last turn, and the user's affect data from the last cycle are buffered in Arny's memory. At stage 4, Arny starts by making a decision about a converging, diverging, or pass action based on this buffered data and the interaction model introduced in Table 2. This action is simultaneously perceived by the user (stage 5) and elicits an emotion in them. The valence value for the elicited emotion in the user is measured by Affectiva and reported to Arny to replace the valence value from the previous cycle in Arny's buffer (Stage 6).

3.4.2 Arny-V1 Collaboration Rules

The collaboration Rules in Arny's model of interaction look into emotion interpretation and details of decision making. Table 2 presents how based on Arny V1 collaboration rules, Arny's awareness of the participant emotional feedback in cycle N-1, and awareness of the participant contribution on the participant turn in cycle N, lead Arny's contribution in cycle N.

Arny V1 model of interaction was designed around a set of assumptions about human emotional feedback to Arny's contribution. Based on the existing literature on emotional feedback (Russell, 1980; Sherer, 2005) negative valence value is considered aligned with any form of emotion that includes a negative value judgment of the process such as confusion, frustration, disappointment, and annoyance. Positive valence value, on the other hand, is in line with emotions that represent a positive value judgment of the

collaborator actions including interest, satisfaction, excitement, convinced, and expectant. For the design of Arny V1 collaboration rules, we hypothesized that if a human's emotion is any of the listed positive emotions, a converging action is more likely to maintain the positive feeling about the collaboration. However, in case of negative valence values a pass action was suggested since a negative valence value is a representation of a negative judgment of the process. Arny V1 assumes that this negative judgment of the process is triggered by a contribution by Arny that distracted the participant from their design goal and so caused negative emotions. Based on Arny V1 model, in such situations Arny must pass on the next action in order to play a milder role in the interaction and reduce the negative emotions about the interaction.

Finally, in cases when the valence value was neutral, maintaining the interaction pattern was considered not beneficial. Based on the interpretation followed by the model, the neutral emotions reflect minimal risk of distracting or annoying the participant in case of a diverging action. Since a diverging action was expected to have minimal risk in such a situation and could encourage the participant's creativity with a shift from the initial idea, diverging actions were suggested by the Arny V1 model in response to neutral valence.

Cycle	Input		Arny's Action
	User's Drawing From Cycle	User's Valence Value in Cycle N-1	
Cycle 1	1	No Previous Reference	Converge to User Drawing in Cycle 1
Cycle N	N	Positive	Converge to User Drawing in Cycle N
		Neutral	Diverge from User Drawing in Cycle N
		Negative	Pass with No Drawing Action

Table 2- Arny Version1 Collaboration Rules

3.5 Summarized Review of Arny V1

In this Chapter, Arny V1 interaction model for co-creative collaboration is discussed. The model uses a turn-taking pattern and incorporates the user's contribution and valence level during each cycle to understand their value judgment of the AI interaction and determine the AI's next contribution. Chapter 4 of this dissertation examines Arny V1's interaction model.

Chapter 4: EXPLORATORY STUDY USING ARNY V1

The exploratory study of Arny V1 was a wizard of Oz study focused on creating a practical emotion-based interaction model and system for creative collaboration with users. To address this objective, the study mainly focused on two high-level goals corresponding to RQ1, and RQ2:

1. Evaluation of the design parameters deployed by Arny V1 based on existing literature and the researcher's conjecture (RQ1)
2. Exploring participants' perception of a co-creative system interacting with them based on emotion detection (RQ2)

The study was performed in the context of an interior design task and narrowed down RQ1 to the following three issues:

A. **Adequacy of emotion detection method:** Is emotion data collection through Affectiva adequate for identifying the emotion stimuli during the co-creative process?

B. **Adequacy of targeted emotion dimension:** Is the valence value captured through Affectiva the proper emotion dimension to effectively interpret the induced emotion(s) in the user during the co-creative process?

C. **Interpretation of the captured emotion:** Is Arny V1's model of collaboration based on diverging and converging actions in response to emotional feedback appropriate for a co-creative setup?

In addition, the study examined the role of Arny's affective model of interaction in shaping users' perceptions of Arny's contributions.

4.1 Methodology

An interior design task was structured as a turn taking collaboration between Arny V1 and the participant to create an office space for a specific scenario. During the study Arny (the wizards) followed Arny V1's model of interaction previously presented in chapter 3 and performed each contribution in accordance with the participants' earlier emotional feedback and the last object added by them to the office space.

The simplicity of a constrained collaboration around an interior design task both in terms of performing the study and analysis was the primary reason for using it instead of a more open-ended task. The presence of restrictions was especially important since the study utilized a wizard of Oz method and the wizards had to engage with the participant actions without a significant delay and in a similar manner to an actual AI.

4.1.1 Experiment Setup

A study script containing study instructions was shared with the participant -who was told will be working with an AI system called Arny- through a simple online collaborative drawing platform that was called Ziteboard. However, there were actually two members of the research team (wizards) sitting in a different room and interacting with the participant. The study setup is presented in Figure 10.

The wizards' interactions with the participant were in accordance with the Arny V1 interaction model and based on the Affectiva's report of participant's emotion which was captured by a webcam on the participant's computer. The wizard team consisted of an Emotion Interpreter Wizard (EIW) and a Drawing Wizard (DW). EIW was in charge of

observing the real-time emotional feedback of the participant on the Affectiva platform and priming DW when to converge, diverge or pass. DW was in charge of contributing to the interior design task according to the primes given by EIW and the last object contributed by the participant. DW had a list of interior design objects to choose from for each contribution. In time for a converging action, DW contributed by drawing an object from the list which was conceptually similar to the user input. For diverging actions, on the other hand, DW drew a relevant object to the task with less conceptual similarity to the input. Arny V1 collaborative study relied on DW to decide on the conceptual similarity or difference between objects on the list and the participant's drawing. Figure 11 represents how EIW and DW collaborate to generate a contribution on behalf of Arny in each cycle.

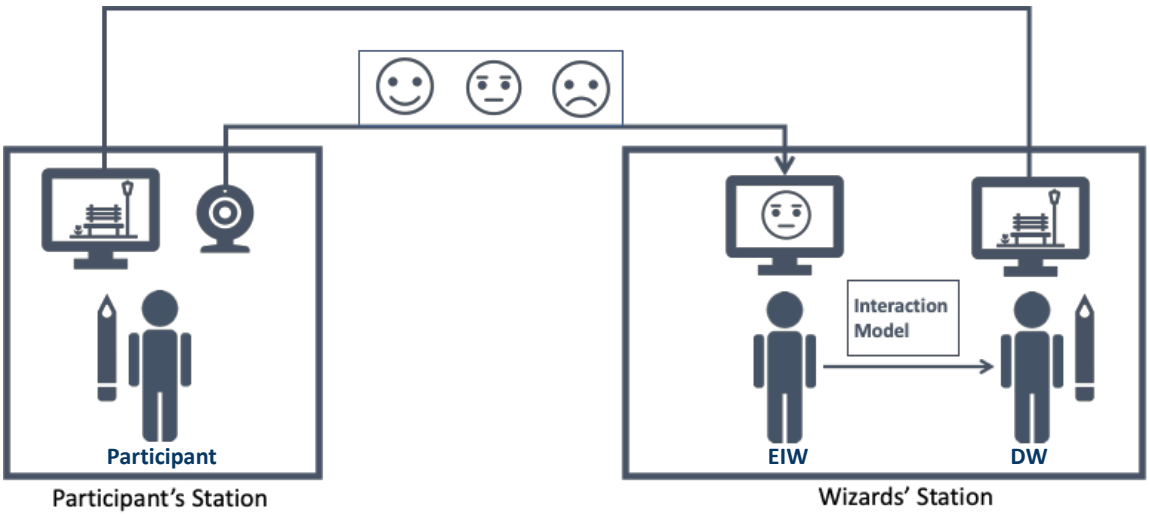


Figure 10- Pilot study setup for testing of Arny V1

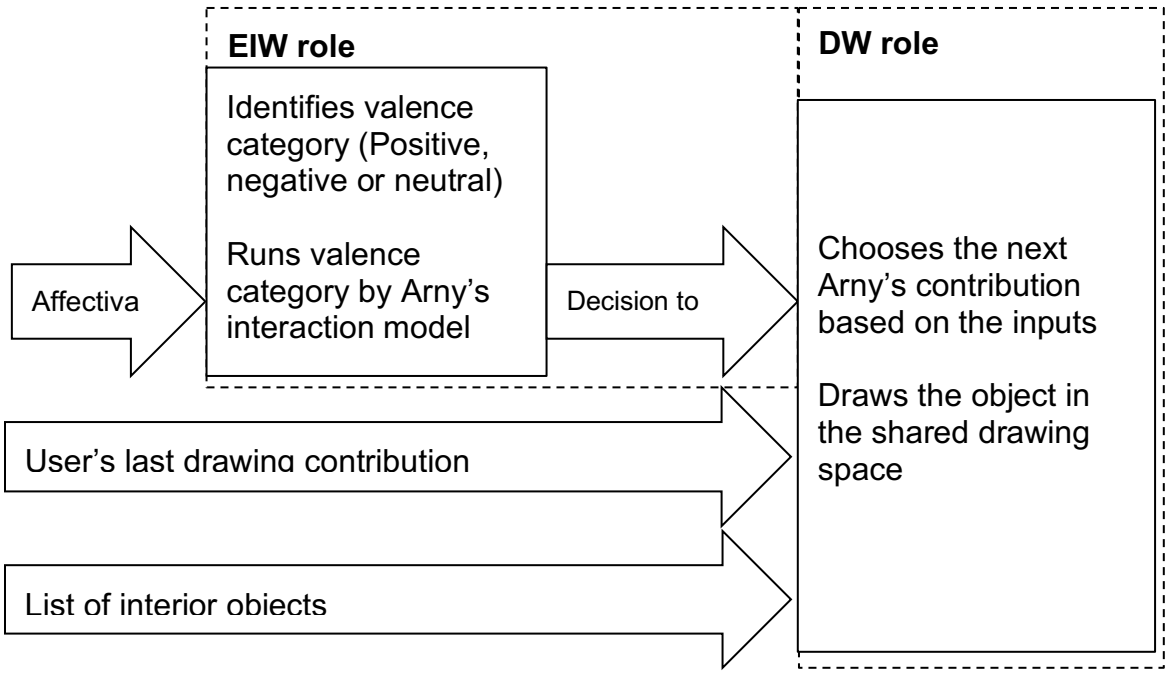


Figure 11- EIW and DW collaboration to generate a contribution on behalf of Arny V1

4.1.2 Recruiting participants

Participants in Study V1 had to meet two key criteria to qualify for participation. 1) have medium or above medium level of design skills, and 2) have no face covering or facial hair

1. ***Level of design skills:*** The main requirement to participate in the study was to have design-related education or skills. This criteria was required since having design related expertise eliminates the occurrence of skill-associated negative feelings such as frustration which could be confounded with the type of affect this study is exploring.

2. ***Facial hair or face covering:*** The second requirement to participate in the study was not to have face cover or lots of facial hair. Participants with facial hair or face cover had to be excluded from the study due to the limitations of the facial expression method of emotion recognition.

In spite of informing the participants about recording their facial expressions during recruitment, pilot runs of the study suggested not to explicitly mention the role of facial expressions before a drawing task. This decision was made after two pilot participants expressed that they were conscious of their facial expression while interacting with the system. One of these participants stated, "I knew the system was watching...so I tried not to show a negative face while drawing".

To fulfill the participation criteria, a line about the requirement of no facial hair and face covering was included in all recruiting announcements. In order to evaluate the design skill criteria and collect some additional demographic information about the participants a pre-screening questionnaire was utilized. The pre-screening questionnaire was built in Google forms and emailed to all study candidates. The form involved demographic questions about respondents age and education background along with screening questions focused on design related training and skills. The details of this questionnaire is available in Appendix 3.

4.1.3 Experiment Procedure

The experiment procedure started with a mini training session where the participant was asked to complete a few incomplete smiley faces together with Arny in a turn-taking

manner. This training session allowed the participant to familiarize themselves with the Ziteboard platform and the turn-taking structure of the collaboration. After the training session, the interior design task description was delivered to the participant by handing them a printed task description, and presenting an incomplete office space on the Ziteboard platform. The incomplete office space was designed in a way that allows the participant and the DW to each contribute more than 10 objects to the space so sufficient volume of data could be collected from each participant. The participants were not informed about the required number of contributions by each collaborator in order to not impact their level of engagement, however the task was stopped for them after completion of 10 collaboration cycles. The study script asked the participant to contribute to the design of an interior office space for the CEO of a tech startup company by adding one object on each turn. Details of this task description are available in Appendix 1. Figure 12 shows a sample of a collaboration between one of the participants and the drawing Wizard.

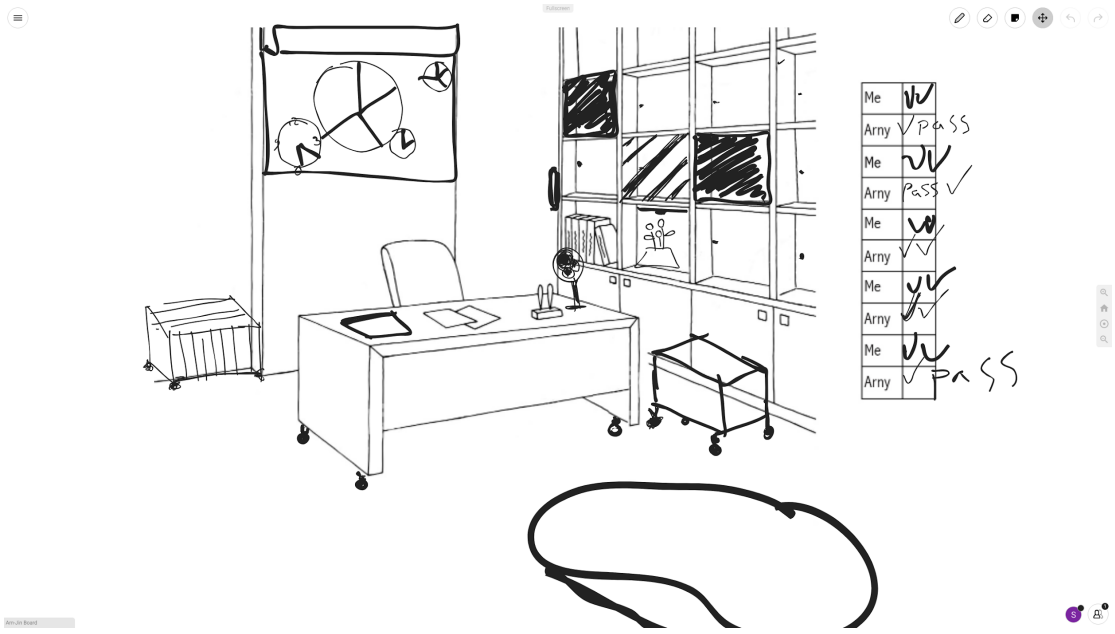


Figure 12- A sample drawing collaboration between a participant and the Wizards on Ziteboard

At the concluding step of the experiment, each user participated in a set of 10 mini-retrospective protocols -one for each cycle of the interior design collaboration- each followed by four interview questions. During the retrospective protocol, the participant was shown a screen recording of the drawing platform and was primed to narrate how they perceived the collaboration. The interview questions were asked at the end of each cycle and were focused on what happened at Arny's turn on that cycle:

1. How did you feel about what Arny did on this turn?
2. Do you think Arny was converging to or diverging from your idea? Did that have an influence on how you felt about it?
3. How do you think what Arny did here influenced your engagement?
4. How did what Arny did influence your next drawing?

4.1.4 Data Collection

In this study five categories of data were recorded and examine to explore the dissertation research questions: 1) screen recording of the collaboration, 2) Affectiva emotion reports, 3) wizards notes during the collaboration, 4) participants' self report during the mini-retrospective protocols, and 5) participants' response to the interview questions.

From this data, the screen recording data, Affectiva emotion reports, wizards' notes and responses to interview question number 1 were used to respond to RQ1. The retrospective protocol data and interview questions 2 to 4 were used to study RQ2. The following subsections describe each data collection category more specifically

1. ***Screen recording of the collaboration:*** Arny's collaboration with each member of Ziteboard was recorded using Affectiva's screen recording functionality. Screen recording through Affectiva makes it possible to review the recorded participants' screen alongside with Affectiva emotion recordings. In addition, Affectiva provides a view of the screen video without displaying the captured emotion diagrams. Figure 13 shows a snapshot of a recorded session displayed alongside with emotion data through Affectiva, and Figure 14 shows a snapshot of a recorded session, not displaying any affect records.

   To aid participants in remembering their actions and describing their emotions more accurately, the retrospective phase of the collaboration featured the screen recording without emotional data. Later, the same screen recordings were used alongside with the emotion data to explore RQ1. The screen recordings as well as the results of the mini-retrospective protocol responses were used to investigate participant perceptions of Arny's actions in order to study RQ2.

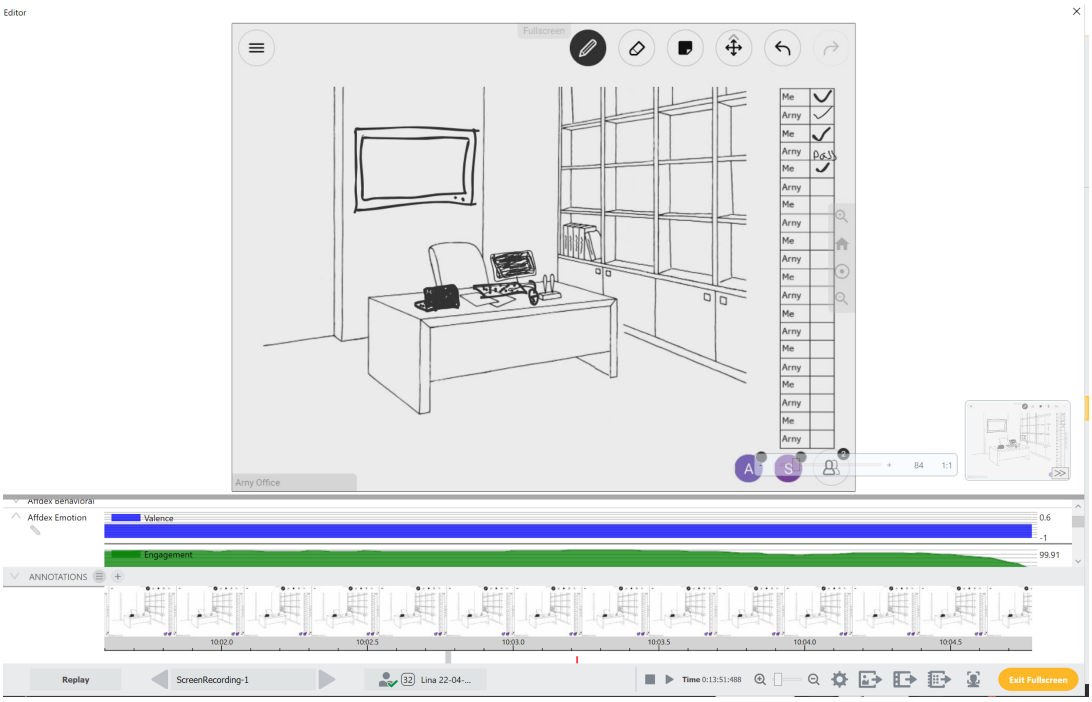


Figure 13- Snapshot of screen recording displayed alongside with emotion data in Affectiva

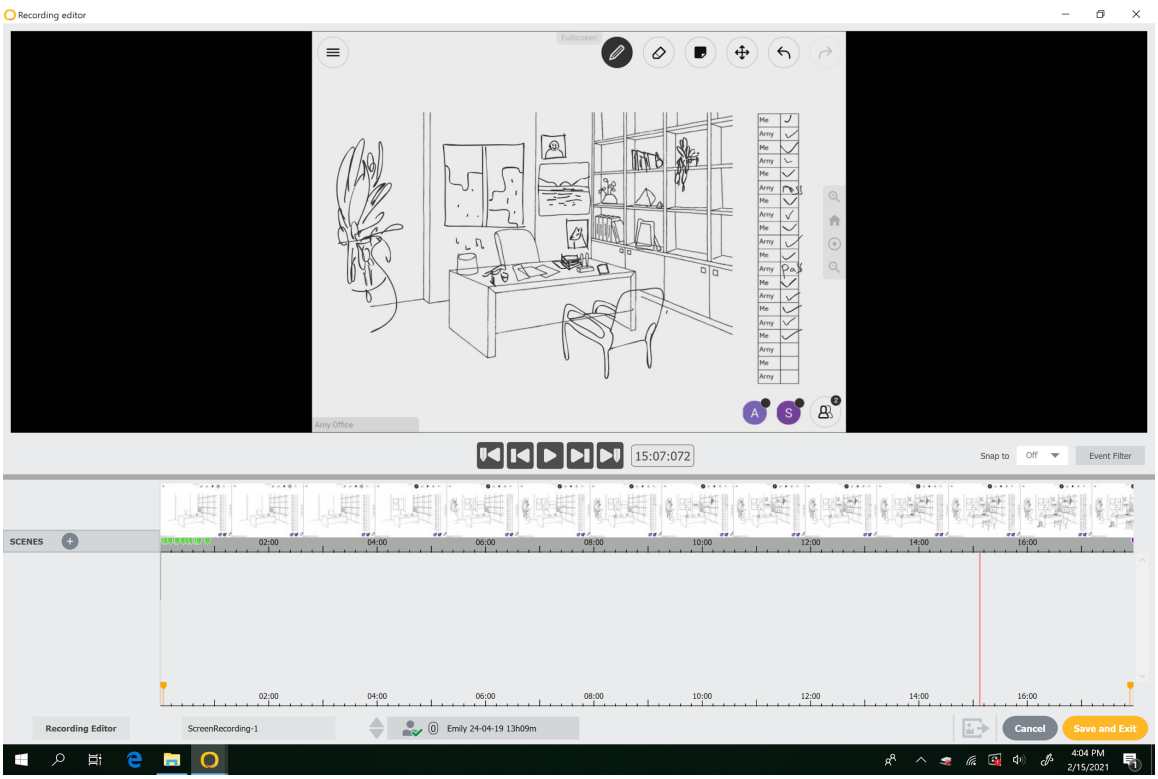


Figure 14- Snapshot of screen recording displayed without emotion data in Affectiva

2. ***Affectiva emotion records*:** Affectiva emotion reports were used in real time to direct the collaborative actions of the wizards during Arny V1. Also, the data from aligned screen and emotion recordings of each task were used for triangulation with the interview and retrospective data during post-study analysis.

   Through Affectiva, researchers can filter emotion dimensions and view a dynamic visualization of them both in realtime and post-study. Affectiva offers various emotional dimensions such as valence, arousal, engagement, joy, and anger. Although the emotion factor used by the Arny V1 interaction model during the study session was limited to valence, this research viewed other emotion factors such as engagement level, anger, joy after the study to investigate adequacy of valence as required by RQ1-B. Later analyses involved triangulating the Affectiva reports with the participant's self-reported feelings to examine any other affect aspects that were referred to by the participants.

3. ***Wizard's notes about the participants actions and emotions:*** EIW tracked wizards due's perception of the participant's actions type and valence category during the study session by registering those values for each turn. To capture these details EIW utilized a table similar to table 3 for each participant. The table is structured so there is a column for each turn within a cycle and each column has a field to record what object was added by the participant or Arny on that turn, if the wizards perceived that to be a converging or diverging action, and if they categorized the Valence value displayed by Affectiva as positive, neutral or negative. This information later

helped us in the analysis phase when working with turns where the participant had more than one valence level in response to an action by Arny.

	Cycle n		*Cycle n+1*	
	Participant	*Arny*	*Participant*	*Arny*
Contributed Object				
Wizard's Perception of Action Type: Diverged, Converged or Passed				
Wizard's perception of Affectiva Valence Category (Pos, neg, Neut)				

Table 3- Structure followed by wizards for taking notes during the Arny V1 study

4. ***Mini-retrospective reports:*** To collect the retrospective reports, each cycle of the collaboration was replayed to the participants with 2x speed and the participant was given time to narrate the cycle. After replaying each cycle, the set of 4 interview questions were repeated to the participants to answer them about that most recent cycle. Participants' responses during this combined retrospective and interviews were 20 minutes on average. These self-reports were audio recorded and transcribed for analysis purposes. Narrative analysis and triangulation with Affectiva data was performed on this data to investigate the research questions. Although the initial intention for collecting the retrospective report was to explore RQ2 on the users' perception of Arny's contributions in association with Arny's model of interaction, this data turned out to contribute to the findings about RQ1 as well.

5. ***Interview responses:*** The four interview questions were repeated to the participants after their retrospective report of each cycle of the collaboration. Participants' answers were audio recorded and transcribed. Answers to interview question 1 were later triangulated with the Affectiva emotion reports. The result of the triangulation analysis was used to examine RQ1 item A in order to determine if the Valence data gathered by Affectiva was sufficient to identify the emotion stimuli. Responses to interview questions 2-4 were used to examine participants' perceptions of the collaboration for RQ2.

4.2 Study Data

The exploratory Study of Arny V1 included 7 participants. The data collection was stopped after 7 participants as it became evident from repeated response patterns that a revised version of Arny would be needed. The data from 1 of the 7 participants had to be disregarded as the person was familiar with the Wizard of Oz method and after a technical issue suspected that he was collaborating with a wizard rather than an AI system.

4.2.1 Study Participants

All the participants were graduate students recruited from UNC Charlotte campus, and had basic or higher design skills according to pre-screening self-reports. Gender distribution for the 6 analyzed participants was 3 males and 3 females.

Participation in the study was voluntary and the participants received a monetary incentive for their participation. The participants were told that the amount of the monetary

incentive will change based on the quality of the collaboration outcome. This was done to encourage them to invest in the task and to engage them in a goal oriented dynamic, similar to common human-human collaborations. However, the incentive amount was the same for all the participants.

4.2.2 Data Analysis & Results Overview

As the study data was collected from multiple sources, the data analysis procedure included varied analysis techniques. An inductive pattern analysis of the emotion data and screen records, a narrative analysis (Holstein & Gubrium, 2011;)of the self-reports during the interviews, and a triangulation of both were performed to explore and interpret the results. Figure 15 presents an overview of the main analysis techniques as well as an overview of the findings from each technique.

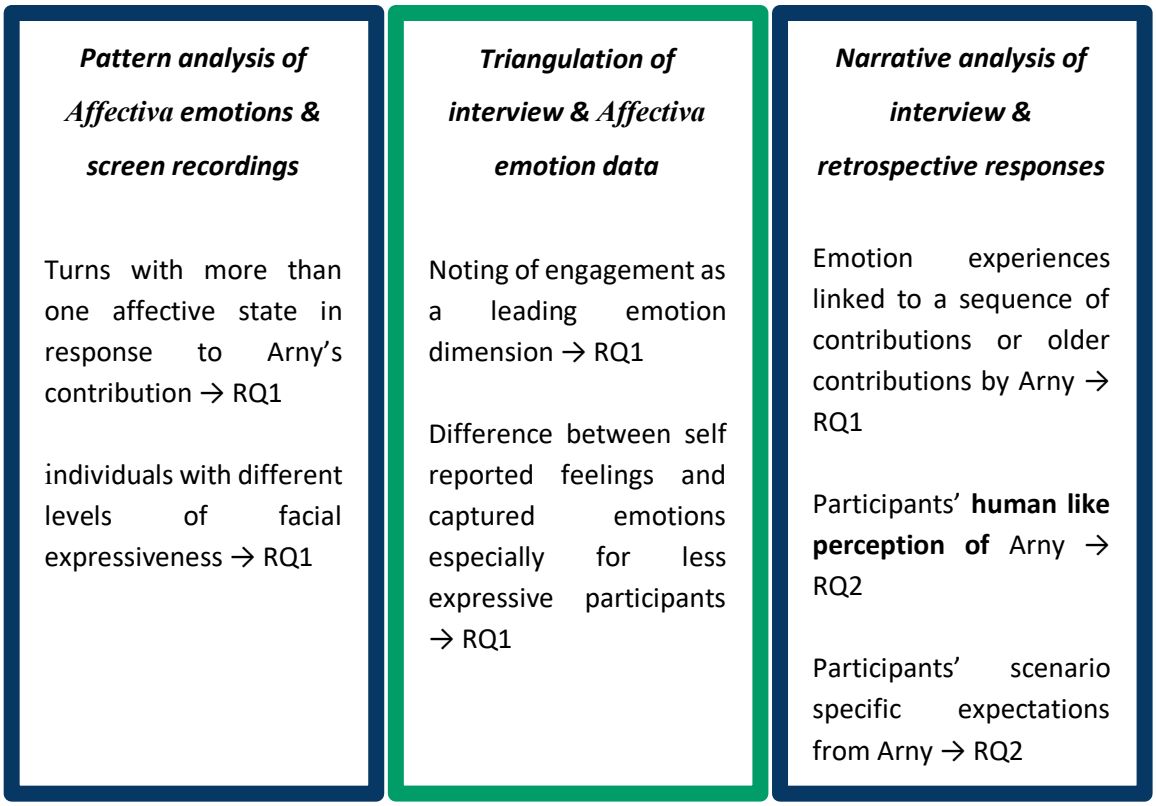


Figure 15- An overview of the data analysis methods and resulted findings from Arny V1 study

Researcher's observation of the Affectiva reports in parallel with the study sessions and pattern analysis of the emotion data and screen recordings identified two findings in relation to RQ1. The first finding was that in some turns the participants experienced more than a single valence value in response to a contribution from Arny. As Arny V1's model of interaction relies on a single valence category for choosing the next contribution this observation suggested a need to revisit Arny V1's feedback collection and interpretation. The second observation related to RQ1 was about the level of expressiveness in different participants. As reflected by the Affectiva emotion reports, facial expressions of different participants were very different in terms of the intensity level of emotional feedback.

Narrative analysis of the self-reports and attitudinal data from the participants revealed that in several cases, participants' emotion experience linked to older contributions from Arny. This finding impacts the design parameters of a co-creative system based on emotional feedback that was explored in response to RQ1. Also, during the same reportings, participants shared their perceptions of Arny and its collaboration actions which was asked in RQ2. Some of these perceptions included **human like intentions and characteristics** of Arny, as well as scenario specific **expectations** from Arny (RQ2)

Finally, the following three insights related to RQ1 emerged from the triangulation of the different data sources. The first finding was that one of the primary emotion dimensions to better interpret a user's stance about the collaboration and respond to it is engagement. The second finding was that the participants' self-reported emotions during the collaboration differed from the Affectiva emotion reports in several cases, which can impact the accuracy of the system following the interaction model. Third, there were turns when reports from Affectiva highlighted multiple affective states on several turns, however, the participants referred to only one or two dominant feelings instead of noting all the different types of emotions that were captured by Affectiva.

The following section will expand on the findings outlined above.

4.3 Interpretation of Findings

The interpretation of the Arny V1 exploratory study results addresses two main questions. RQ1 is addressed in section 4.3.1 by examining Arny's approach for detecting,

targeting, and interpreting emotion. Section 4.3.2 addresses RQ2 on participants' perception of Arny V1's interaction model and the collaboration experience.

4.3.1 Adequacy of Emotion Recognition Method

RQ1 was narrowed down to three design aspects at the beginning of chapter 4: a) adequacy of Arny's emotion collection method, b) adequacy of the targeted emotion parameter (valence) for interpreting induced emotions in the users, c) appropriateness of how Arny's model of interaction interprets positive, neutral and negative valence values.

In Arny V1's study, multiple issues emerged regarding the adequacy of the emotion collection method. The first issue was that Arny V1 failed to accurately detect emotions in participants with less expressive faces. The second challenge was in some cases participants experienced more than one valence category (positive, neutral, negative) in response to a single action by Arny. Third, in a few cases triangulation of Affectiva and retrospective data revealed that participants had experienced emotions related to an Arny's action from a previous cycle.

Concerning the sufficiency of valence and the adequacy of the interpretations of positive, neutral, and negative categories of valence values, retrospective reports revealed that in some instances Arny V1's interpretation of negative and neutral valence values differed from participant's feelings about collaboration. Such differences were pointed out by participants in retrospective reports of the turns when the participant had lower engagement as they were running out of ideas and expected a creative idea from Arny to help them move forward. Triangulation of these retrospective data with Affectiva reports

suggested that including engagement values may contribute to the accuracy of Arny's interpretations. The following subsections discuss these findings in depth.

Stoic vs expressive faces and affect reports

Arny's collaboration with the 6 study participants resulted in a total of 120 drawing objects from which 60 were contributed by Arny. The following significant differences were found when comparing the participants' self-reports of affects with Affectiva's reports for the 60 collaboration turns:

Positive emotions: Based on the Affectiva recordings participants' affect was positive in 14 out of the 60 cases, while participants self-reported positive emotions in 32 out of the 60 cases.

Neutral emotions: Based on Affectiva reports, participants experienced neutral emotions in response to 26 Arny turns. However, participants self-reported being neutral about Arny's contributions in 15 turns.

Negative Emotions: Affectiva reported negative emotions in nine cases while the interviews reported eight cases.

***Multiple emotions*:** Affectiva reported the participant experiencing multiple emotions in response to 11 out of the 60 Arny turns. In interviews, participants reported experiencing more than one emotion five times out of 60. More details on multi-emotion turns will be discussed later in this chapter.

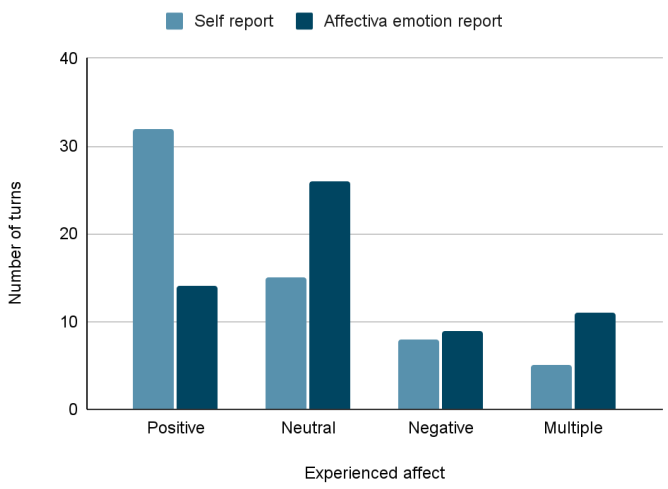


Figure 16- Overview of emotion reports from the two emotion report sources of Arny V1study, Affectiva, and participant interview responses

Triangulation results revealed that 2 of the study participants had less expressive faces compared to the other four (Figure 16 & Figure 17). In 13 out of the 20 turns for these participants, the affect reported by Affectiva was more negative than their self-reported feelings, i.e. 1 case the valence recorded as negative by Affectiva was reported as neutral by the participant and in 12 cases the valence was reported as neutral by Affectiva but reported as positive by the participants.

Further explorations of the emotion recognition literature suggested that demographic factors such as race, ethnicity, and gender can impact facial expressions for different emotions as well as the expression intensity (Xu et al, 2020; Dailey et al, 2010). This finding led us to using consistent demography for the next Arny evaluations. A future resolution to this challenge which is outside the scope of this dissertation is to deploy a calibration feature that is based on the results of an introductory task performed by each user.

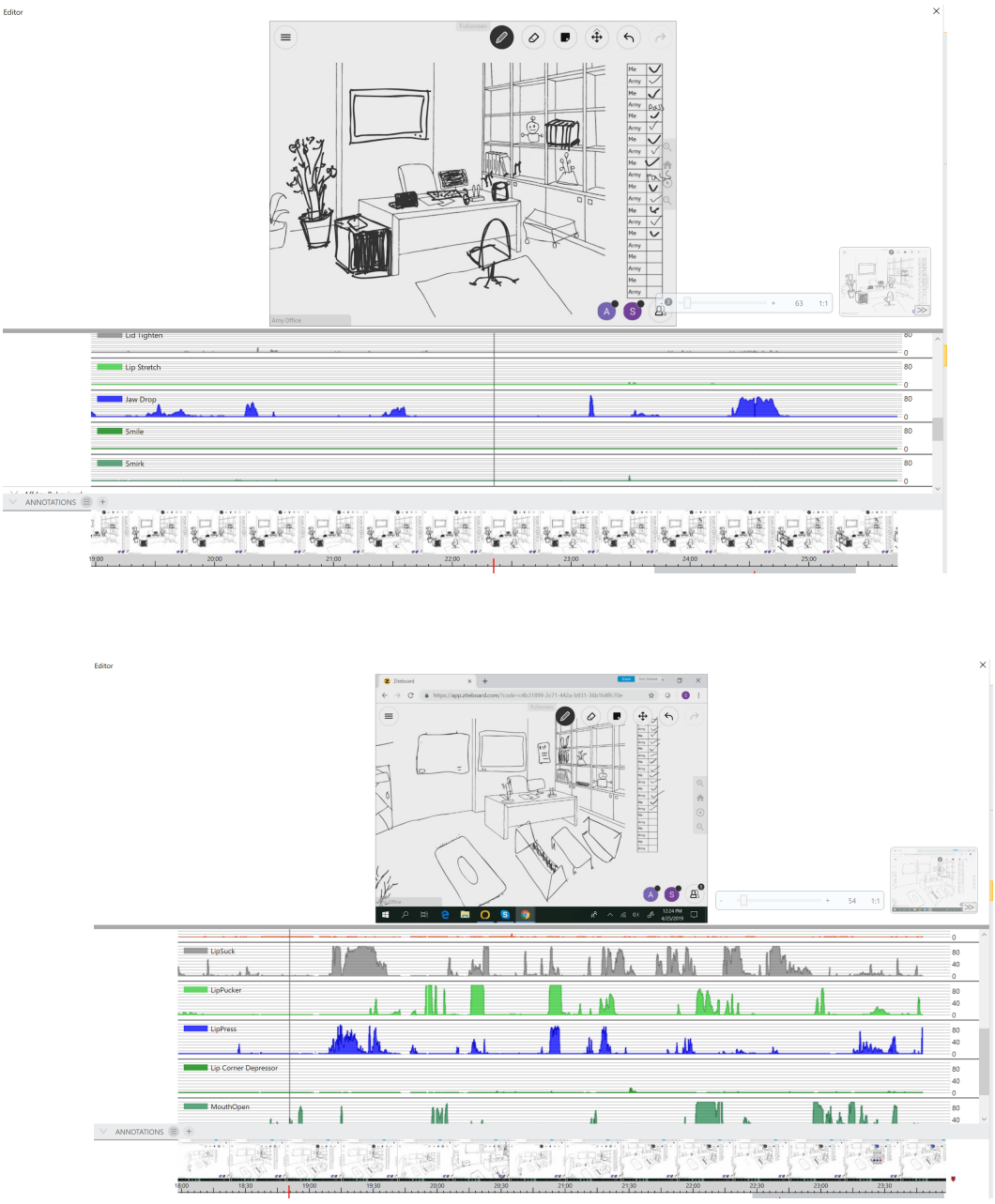


Figure 17- Snapshot of participant facial expressions for a participant with a stoic face (top) compared to a participant with a more expressive face (bottom)

Multiple-emotion turns

Unlike the expectations to observe one emotion in response to each event, the data reflected cases when Arny's actions elicited more than one emotion during one turn. This

dissertation refers to such turns as multi-emotion turns. Figures 19 & 20 present multi-emotion turns for one of the Arny V1 study participants.

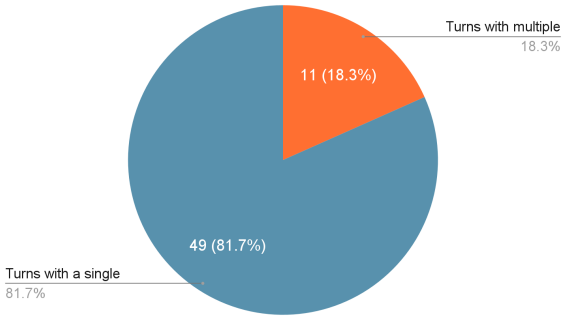


Figure 18- Frequency of multiple emotion turns throughout Arny V1 study with a total of 60 Arny turns performed in collaboration with 6 participants

The Arny V1model of interaction requires the EIW to record a single valence level for each turn in order to decide on Arny's next actions, therefore EIW had to report his perception of the most dominant valence level for each turn while reading Affectiva data in real-time. The EIW's perception of dominant valence levels were later compared to participants' self-reported feelings during interviews and retrospective protocols. Affectiva reported 11 multi-emotion turns during Arny V1 study (Figure 18). In 6 out of these 11 cases, the participants reported only one emotion which was the same emotion as the dominant emotion perceived by the EIW. In the other 5 cases however, the participants also noted more than one single dominant feeling for those turns. The mismatch between the two reports can be explained by people's tendency to report overall feelings rather than the details of the momentary emotions (Prinz, 2005)

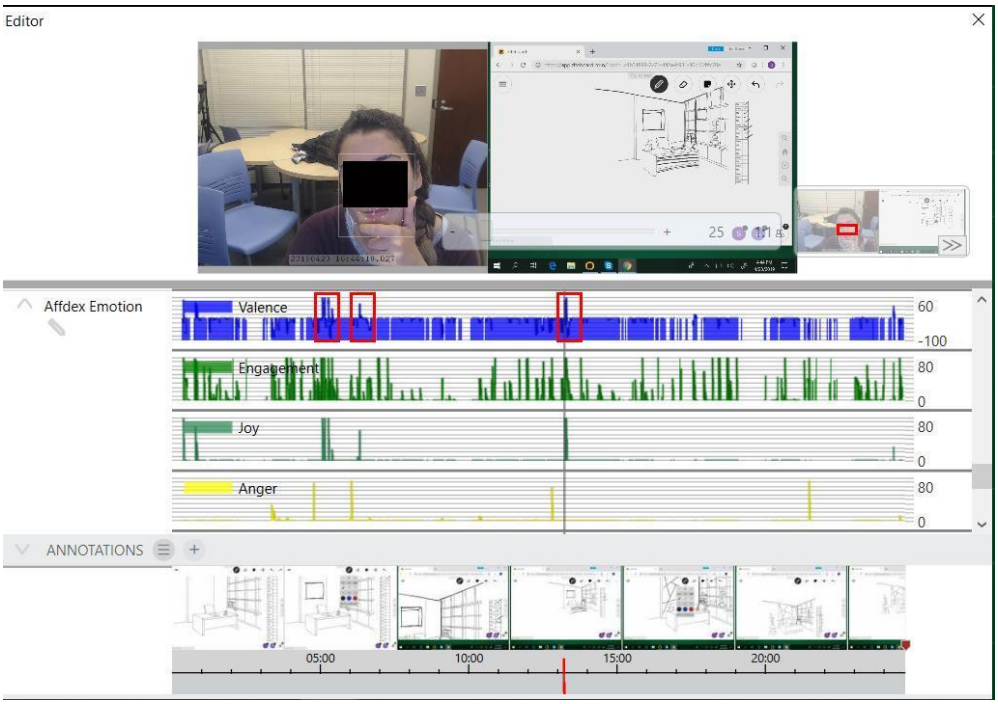


Figure 19- A sample participant's affect changes throughout the design task- Areas marked with red present multiple valence values during one turn

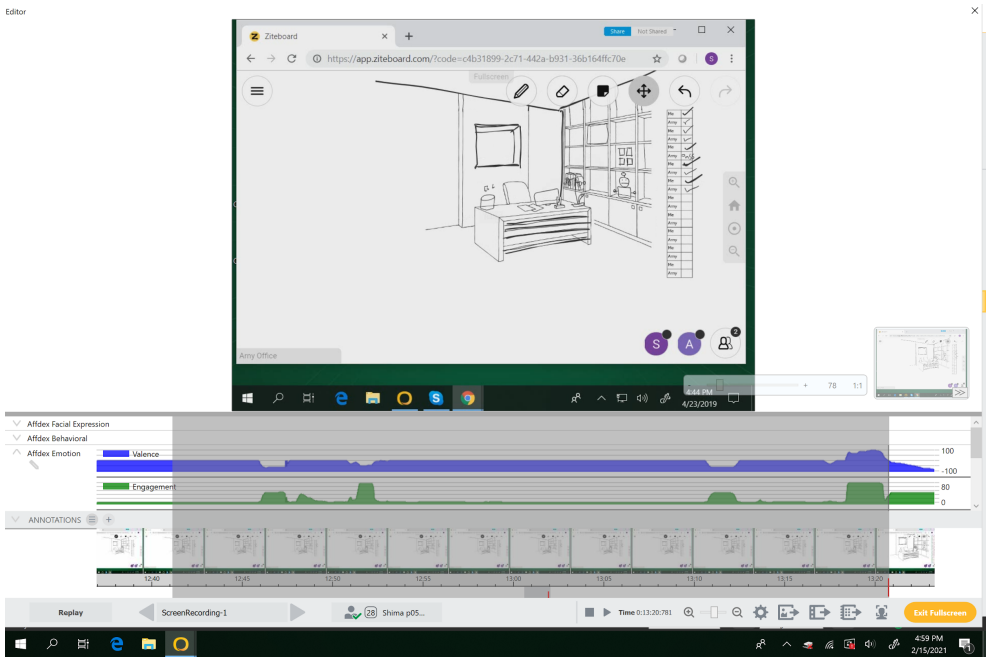


Figure 20- Zoomed in view of the participant's affect changes on one Arny turn marked by the gray box

The self-reports of participants revealed two main causes for multiple emotion turns while exploring the underlying reasons, 1) the participant being unsure or confused about how to interpret a contribution resulted in a change of judgment as that contribution was made, or 2) the participant liked some aspects of a contribution and disliked some other aspects of it:

> *"At first I* didn't *know what it was; but then I realized it was a robot. It was really cute"*
>
> *"The system added flowers. I liked the idea but not the position"*
>
> *"I didn't like the shape of the chair. Though I like the idea of chair.*"

Flash-back emotions

Another pattern that emerged from the Narrative analysis of the retrospective protocol was what this report refers to as the flashback pattern. Flash-back emotions are user emotional feedback in relation to an Arny's actions in a previous cycle rather than the most recent one. In 4 turns during the retrospective protocol, the participants reported emotions which were completely or partially relevant to a previous contribution from Arny rather than the latest contribution which was in review.

> "I wasn't really excited about this turn. I felt like it (Arny) was trying to pulse-back to the functional stuff we were doing at the beginning, which I didn't enjoy at the beginning, so I wasn't really super happy"

For future research, eye-tracking can be used as a possible method to identify flash-back emotions where a contribution other than the most recent object is triggering an affect.

Future versions of affect based co-creative systems can deploy eye-tracking as well as more complex collaboration rules to identify and respond to such events.

Association between affect and expectations

The Arny V1 study revealed that in some cases participant expectations, either met or unmet, determined the emotional response to Arny's actions, rather than the what was the type of the performed action. Based on the observations, the participants' positive responses were not correlated with specific action types- converging, diverging or pass. However, interview results revealed that positive or negative feelings were correlated with how well the participants' expectations from Arny were met. i.e., the participants had positive feelings about a converging action when they perceived Arny as building on their idea, and disliked a converging action when they expected an inspiration that they did not receive. Similarly, the participants' liked a diverging action when they perceived it as an inspiration "just on time" and disliked it when they perceived it as distraction instead of the "cooperation" they were expecting to see. For the pass actions also, participants stated liking them when they had a plan for the future of the task and felt more in control, but disliked them when they were unsure of what to contribute next and their expectation for help was not met. For example, the data revealed that in 3 out of the 9 captured negative valence, the negative emotion was a result of Arny passing when the participants were looking forward to an inspiring contribution.

"As I said before, I didn't have much *ideas here. I was waiting for some ideas. (Arny passed) It was annoying."*

Similar to that, participants reported negative emotions during those converging actions when Arny's contribution lacked the creativity level expected by them.

While only valence values cannot predict the participants' creativity and inspiration expectations, revisiting the Affectiva emotion records in accordance with the retrospective reports suggested the future interaction model can consider the participant's engagement level in addition to the valence value in order to detect their inspiration expectations.

4.3.2 Participant's perception of Arny's actions

Participants' interpretation of Arny's intention behind each contribution was one of the interesting patterns in the retrospective protocols. Although DW's actions were all performed based on Arny's model of interaction and not the wizards' choices, participants had a human-like interpretation of these actions. Participants personified Arny as a collaborator who was: 1) Understanding/not-understanding them 2) predicting them 3) Considering their ideas and building on them, and 4) Inspiring them and surprising them by bringing new ideas to the table. This was especially interesting when participants interpreted Arny's action in relation to its capability to understand their stance about the collaboration: *"Arny understood that I didn't like what it did"*, *"It was giving me space to develop my idea," "I liked that it cared about what I did,"* and "*This was kinda freaky...I was thinking of drawing a circle here and then Arny did literary what I was planning to draw...so that was great."* These types of statements about users' perception of Arny's actions confirmed that the interaction model used for collaboration was successful in interactional sensemaking of the process and creating a human-like interaction dynamics pattern between the user and Arny. Furthermore, in some other cases, participants tied

Arny's pass action with it failing to understand their contribution. This was while the interaction model followed by Arny was not at all considering the participant's drawing when it came to pass actions: *"I think it was struggling with understanding my last contribution because it was different from the rest of the drawing."*

Chapter 5: ARNY V2, AI-BASED ENGAGEMENT EXPECTATIONS INTERACTION

Arny V2 is an enhanced version of Arny's interaction model for a co-creative system based on emotion feedback. Arny V2 incorporates improvements in the following areas, based on the results of Arny V1 study:

1. Incorporating engagement level to understand users' expectation from AI
2. Revised collaboration rules based on valence and engagement inputs
3. Offering an instruction to target one affective state in multi-emotion turns
4. Transitioning from Woz approach to a partially developed system

This chapter will review details of the above revisions reflected in Arny V2's interaction model. Chapter 6 will discuss a case study of this revised design.

5.1 Arny V2 Interaction Model

Similar to Arny V1, the interaction model in Arny V2, follows a turn taking strategy. Arny V2 interaction model is illustrated in Figure 21. This new interaction model selects the next contribution from the AI agent based on three inputs: 1) a memory of the collaborator's valence in the previous contribution cycle that allows Arny to perceive the user's value judgment of that last contribution, 2) the user's engagement level in the beginning of the current cycle that allows Arny to predict user's expectation from Arny in the current cycle, and 3) the user's latest contribution to the task in the current cycle. The first input allows Arny to evaluate how satisfied the user is with the previous interaction by the agent and if the same type of contribution should be followed or the pattern has to change. The second input, allows Arny to know if the user expects Arny's assistance for

discovery of new ideas or if they can continue without a major contribution from Arny. This input was added to the Arny's latest iteration after participants testing the previous interaction model referred to their expectations of help from Arny as a trigger for parts of their emotions. Finally, the user's drawing contribution is the third input, to shape the functional sensemaking around the artifact.

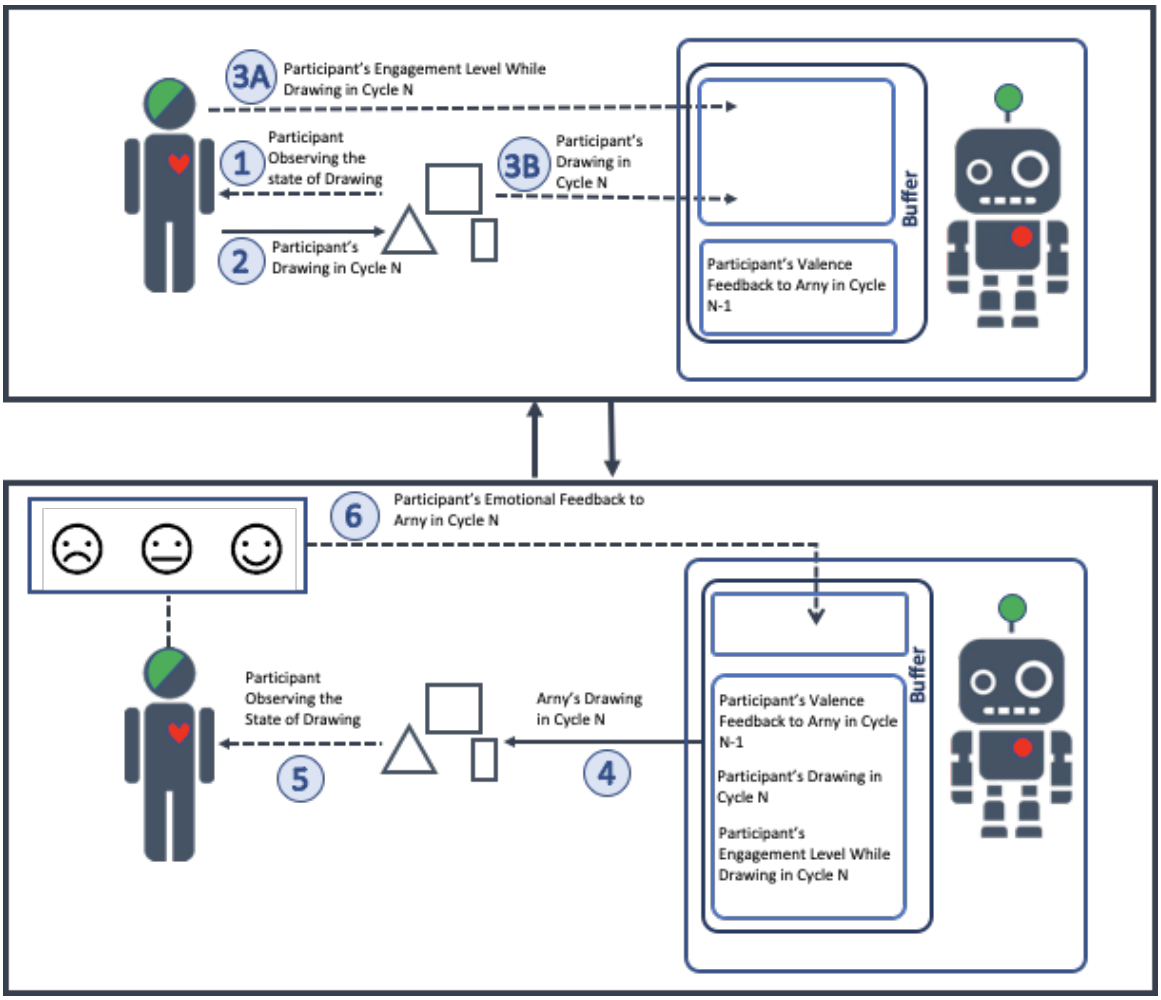


Figure 21- Arny V2 Interaction Cycle- Participant Turn (top) and Arny Turn (bottom)

Interpretation of these three inputs and decision making suggested by Arny V2's collaboration rules are formed based on Arny V1 exploratory study. Arny V1's study learnings inspired Arny V2 collaboration rules to have a strong reliance on engagement

both for interpretation of the valence value and understanding of the participant expectation after they pass their own turn. The study suggested that the main two triggers for negative emotions in a user are the AI causing an unpleasant distraction from the user's idea or the AI not meeting the user's expectation for help in ideation. Arny V2 distinguishes these two triggers based on the flow of user engagement and responds to them differently. In Arny V2 collaboration rules, negative valence value is considered aligned with any form of emotion that includes a negative value judgment of the process such as confusion, frustration, disappointment, boredom and annoyance. Positive valence value, on the other hand, is in line with emotions that represent a positive value judgment of the collaborator actions including interest, satisfaction, excitement, convinced, and expectant. Arny V2's collaboration rules are shown in Table 4. In conditions where the user's emotion is positive and the engagement level does not reflect a desire for creative ideations from Arny, a converging action is more likely to maintain the positive feeling about the collaboration. In the case when the positive valence is followed by low engagement from the participant, there is a risk of boredom or distraction from the task and so a diverging action is presented by Arny.

Cycle	Input			Arny's Action
	User's Drawing From Cycle	User's Engagement Level in the beginning of Cycle N	User's Valence Value in Cycle N-1	
1	1	Not referred to for the first cycle	No Previous Reference	Converge to User Drawing in Cycle 1
N	N	Medium to High	Positive	Converge to User Drawing in Cycle N
			Neutral	Converge to User Drawing in Cycle N
			Negative	Pass with No Drawing Action
		Low	Positive	Diverge from User Drawing in Cycle N
			Neutral	Diverge from User Drawing in Cycle N
			Negative	Diverge from User Drawing in Cycle N

Table 4- Emotion interpretation and decision making in Arny V2

5.2 Arny V2 System Components

Unlike the Arny V1 model of interaction which was fully deployed as a Wizard of Oz concept, Arny V2 is a combination of Affectiva emotion recognition platform as the emotion detection component that captures user's emotional state during the collaboration, and an AI model for selecting sketches to contribute to the current design. The user experience is an interior design task designed on Microsoft PowerPoint to collect the user's contribution and present Arny's contribution in response to the user. For this system design, the Teamviewer remote access and control software and the wizard team were used to facilitate the communication between the three main system components. Figure 22 demonstrates Arny V2 components and the interaction between them.

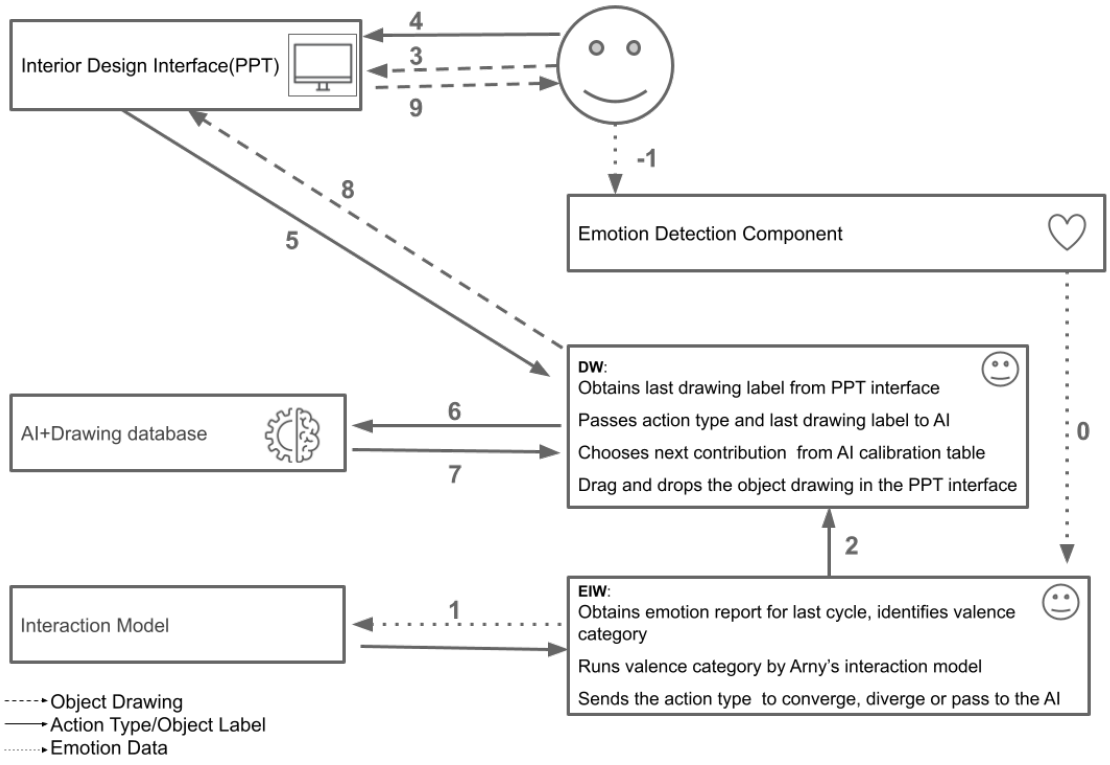


Figure 22- Arny V2 System Components

5.2.1 Emotion Detection Component

Arny V2 uses real time Affectiva reports of engagement and valence. Similar to emotion detection in Arny V1, Affectiva software and a webcam are used for emotion detection and real-time values are calculated from the video of the user captured through a webcam on the user's computer.

As in Arny V1 study, EIW categorizes the observed valence values from each turn as positive, neutral, or negative. However, to stay consistent on the multi-emotion turns, EIW considers the latest valence response from the participant on the Arny's turn. The rationale for this approach is while the participant can experience multiple emotions in

reaction to a single contribution from Arny, their latest affect represents their last position in response to Arny's contribution.

For engagement dimension, EIW observes the real-time input from Affectiva and categorizes the numeric engagement value as either low engagement or medium/high engagement. Unlike turns with multiple valence values, in turns with changing engagement levels, EIW's categorization of engagement is based on the early engagement in that turn. This is since the participant's engagement is a reflection of their desire for creative ideation from Arny and is measured in the beginning of their turns as they start to choose their next contribution.

5.2.2 Interior Design Interface

The user experience provided to the participants to take part in this study is a simple interior design interface created in Microsoft PowerPoint. Three main reasons led this research to choose Microsoft PowerPoint instead of developing a new interface or using more complex online drawing platforms. First, PowerPoint drawing has a very simple interface with a pen tool and an eraser tool on the top left corner of the screen. As previously explained in the design considerations section, it is critical for Arny to reduce the emotional feedback that is potentially triggered due to the complexity of the platform. Second, PowerPoint allows the users to use a stylus pen to draw on the screen and does not limit them to mouse interaction. This is of particular importance to this study since study pilots showed that limiting the participants to only draw with a mouse caused them to feel a high level of frustration. As the frustration elicited due to interaction challenges with a mouse could be confused with the emotions elicited by the collaborator, it is important to

provide the participants with a stylus to eliminate such potential emotions from the study calculations. Lastly, the purpose of a study of this kind is to investigate a new avenue for understanding human-AI creative collaboration. Using a newly designed interface developed with resources available for such study could however bias the user's perception and feedback about the whole idea. In order to focus on emotion-aware collaboration and eliminate the impact of Arny's interface on affect, a platform developed with commercial resources was utilized. This direction allowed to reduce the bias a simplistic interface designed and deployed by the research resources could introduce to the study.

Figure 23 is a presentation of the PowerPoint frontend for the study. This interface includes a drawing area where the participant and Arny draw a new object on their turns in each cycle, a text area in which the participant tells Arny what they drew in each turn and an interior background to which the participant moves and places each new object drawn by them or by Arny. Microsoft PowerPoint is chosen for the purpose due to its simplicity and familiarity for the average participants that allow the study to reduce emotions triggered by confusion about the interface or frustrations that could be triggered by unexpected technical issues possibly occurring in a newly developed interface.

For the purpose of this study it was decided to not rely on feature extraction algorithms to identify the object contributed by the participants and instead require the participants to label what they contribute on each cycle.This decision allowed for the study to be free of bias introduced by inaccurate object recognition and focus on Arny's choice of contribution based on the emotional feedback. Similarly, the decision to rely on the participant for resizing and positioning the object on the background was made to reduce

the decision making variables and only focus on what was impacted by Arny's emotion awareness.

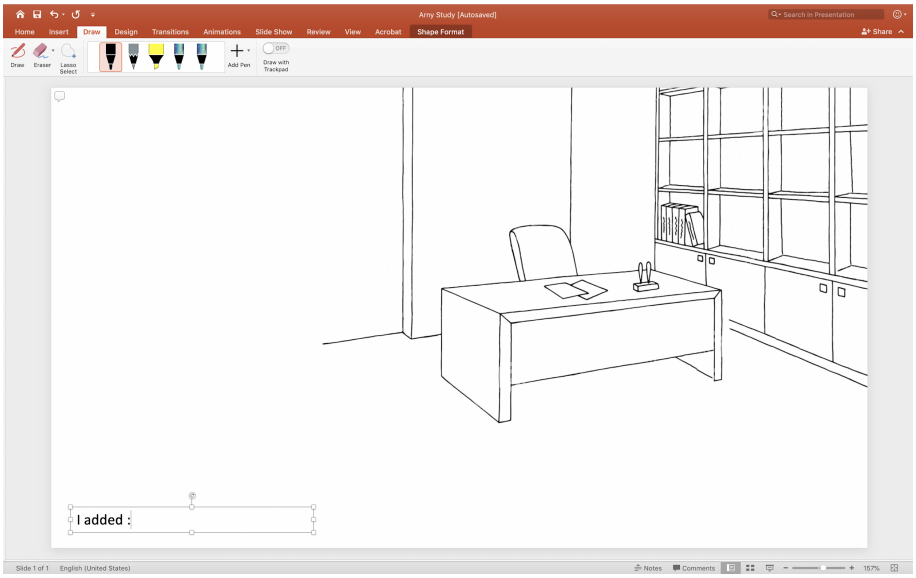


Figure 23- Arny's Design interface in Microsoft PowerPoint

5.2.3. AI Component

The AI component of Arny V2 determines Arny's contribution to the collaboration based on the user's drawing contribution and Wizard's prime of the action type. As previously mentioned in Arny's interaction model, the collaboration between Arny and the user is shaped in the form of interaction cycles. Each cycle starts with a contribution from the user followed by a converging or diverging contribution, or pass action by Arny V2 that is decided on based on tracking the user's previous valence and engagement levels. On each cycle, the AI component is primed on the type of action to be taken and translates this action to a sketch to be contributed in response to the user's last action. To do so, Arny selects from a database of sketches relevant to interior design.

The data set of sketches includes 102 labels and corresponding sketches partly generated by a researcher of this study and partly from the human sketch dataset (Eitz et al, 2012). When participants draw a sketch, Arny takes the label of the sketch and uses a word embedding model (Mikolov et al, 2013) to prepare a list of converging and diverging labels from the 102 objects. Word2vec model is trained on a Wikipedia dataset and represents objects as vectors and puts similar objects together in the vector space. Cosine similarity scores capture the angle between two vectors. If cosine similarity is 0, the angle between the two vectors is 90 degrees, which means the vectors are not similar. Conversely, when the cosine similarity is 1, the angle between two vectors is 0 degree, which means two vectors are pointing in the same direction, thus similar to each other. The Gensim Python library (Rehurek & Sojka, 2010) is used to calculate the cosine similarity score of two labels and consider the score as the indicator of divergence and convergence level. Two labels diverge from one another when their cosine similarity score is less, and converge to one another when their cosine similarity is more. Arny calculates the cosine similarity scores between the label of participant's sketch and the labels of 102 objects and presents a ranked list of the top 10 labels with higher similarity scores to the study wizards.

Wizards in the study ensure that Arny selects the appropriate labels for generating sketches in the collaboration. Wizards are involved in this decision based on the observation that sometimes cosine similarity scores in the word2vec model do not reflect the relevance to the collaborative environment. The reason is that the model is trained on the Wikipedia dataset, but the study focuses on the interior design environment. Before onboarding the wizards as a resolution to this challenge, the word embedding model trained on Google news was also investigated. After examining different combinations of cosine

similarity score ranges and models, it was determined that the top 10 cosine similarity scores in the word embedding model trained on Wikipedia maintained the relevance and did not diverge too much that compromised the context compared to the model trained in Google news.

Arny selects the top 3 of that 10 labels list as converging labels and bottom 3 of the list as the diverging labels. Table 5 shows Arny's calibration table for 'laptop' and 'picture' where Arny identifies that 'computer', 'cell phone', and 'iPod' labels are converging to 'laptop', whereas 'headphones', 'calculator', 'binoculars' labels are diverging from 'laptop'. Wizards consider both participants' emotional response and the calibration table generated from decision making AI to facilitate Arny's next action type: diverge, converge, or pass (Abdellahi et al, 2020).

User's sketch label	Ranked List of top 10 high cosine similarity scores									
	Converging labels							Diverging labels		
laptop	computer	cell phone	ipod	keyboard	phone	printer	camera	Head phones	calculator	binoculars
picture	camera	poster	painting	tv	book	window	face	tape	computer	table

Table 5- Arny's calibration table for convergence and divergence

5.2.4 Wizards

The Wizards' role in current development of Arny V2 interaction model for co-creative collaboration is to facilitate the communication between the other components and supervise the object selection by the AI model. Similar to Arny V1, the wizard team consists of an Emotion Interpreter Wizard (EIW) and a Drawing Wizard (DW). The EIW

is in charge of observing the real-time emotional feedback of the participant on the Affectiva platform and priming the DW when to converge, diverge or pass. In this deployment, the interaction between the emotion detection software and the AI is handled via a wizard due to the difficulty of automatically correlating trigger events and the emotion response, which is beyond the scope of this study.

Arny V2 DW is in charge of contributing to the interior design task according to the input from the EIW and the last object contributed by the participant. This wizard passes the participant's label for their last contribution to the AI model and receives a list of 10 relevant sketches in return. When converging, DW chooses the most relevant sketch from the top 3 objects in this list and copies the selected sketch from the drawing dataset to the interior design interface. In the case of diverging action, DW chooses the most diverging sketch from the last 3 items in the list of top 10 sketches and copies the selected sketch to the interior design interface.

5.3 Summarized Review of Arny V2

In this chapter, Arny V2 interaction model for co-creative collaboration was discussed. This interaction model is a revision of Arny V1 interaction model informed by the Arny V1 evaluation study results described in chapter 4. Arny V2 interaction model incorporating user’s valence and engagement levels to understand their value judgment of the AI interaction as well as their future expectation from AI. An evaluation study of Arny V2’s interaction model and deployment is presented in chapter 6 of this dissertation.

Chapter 6: EVALUATION STUDY USING ARNY V2

The evaluation study for Arny V2 is a case study aiming to

1. Explores RQ1 and RQ2 for the revised emotion recognition and interaction model deployed by Arny V2
2. Explore RQ3 through comparing Arny V2 emotion-based collaboration versus a control condition that doesn't account for emotional feedback.

Arny V2 evaluation study is structured as a controlled case study around two interior design tasks very similar to Arny V1 study. The following chapter presents the study methodology, the study findings, and the conclusions.

6.1 Methodology

Arny V2 case study used the same interior design layout as Arny V1 study. The study had two counterbalanced conditions with the same layout and slightly different task instructions. Controlled condition asked the participant to contribute to an office design for a female faculty and treatment condition asked the participant to contribute to the office design for a female startup manager. The treatment condition Arny V2's model of interaction based on emotional feedback from the participant to collaborate and decide on convergence and divergence. However, the decision on when to converge or diverge was random in the controlled condition. To provide the participants with images that converge or diverge with their ideas, the wizards used the trained word embedding model (discussed in section 5.2.3) to choose the sketches in both conditions.

6.1.1 Study Setup

The study happened in a lab setup with two separate rooms. The participant sat alone in one of the two rooms and the two Wizards worked from the second room. A microphone in the participant's office allowed the researcher to hear potential verbal feedback from the participant. A webcam on the participant's computer was used to collect facial expression input to be analysed by the Affectiva software. The participant was presented with the initial office layout shown in Figure 23 in Microsoft PowerPoint i.e the interior design interface, which was shared with the wizards through TeamViewer along with the participant's Affectiva emotion readings. The Wizards passed the input from the participant to the AI Model and back. The computer used by the participant was equipped with a stylus pen, and webcam that was used by Affectiva software which was running in the background on the participant's computer. The Wizards' lab was equipped with two wide monitors, one for real-time presentation of Affectiva emotion reports to EIW, and access the AI I/O and one for the DW to access the Interior design interface shared through TeamViewer as well as the drawing database to drag and drop the objects into the Interior design interface. The lab setup for Arny V2 study is mapped in Figure 24.

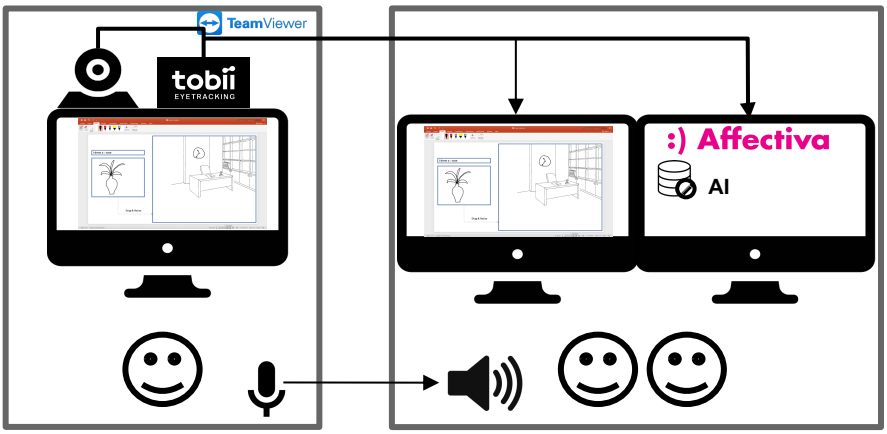


Figure 24- Lab Setup for the study

6.1.2 Recruiting participants

For the Arny V2 study, in addition to the Arny V1 study criteria which included the participants' design skills, facial hair, and face coverings, demographic criteria were also added to maintain consistency across the participants. The criteria was added after the observations from Arny V1, as well as deeper literature reviews (Xu et al, 2020; Dailey et al, 2010) reflecting the influence of demographic factors such as ethnicity and race on facial expressions. Participation also required an expressive face, as developing an expression calibration feature to address less expressive faces is currently outside the scope of this dissertation.

To identify qualified participants, Google Forms was used to conduct a pre-screening evaluation. The pre-screening asked the candidates about their demographic characteristics and basic design skills. Visual appearance, facial expressiveness, race, ethnicity and educational background were other main criteria for recruiting the evaluation study participants. Through this pre-screening, CS undergraduate candidates who reported that they did not have facial hair and considered themselves to be facially expressive cocasian Americans with medium or above medium drawing skills were recruited. The study targets Caucasian Americans because the data used to train existing emotion recognition models largely comes from this demographic (Xu et al, 2020), which can provide more accurate results for the purposes of Arny's evaluation. CS undergraduates were targeted since the research team could access them more easily than students with other educational backgrounds.

6.1.3 Evaluation Procedure

Each study session started with an introduction about the study, Arny, and the interior drawing interface. After the introduction, the participant received a task description for one of the two tasks (controlled or treatment). The participant and Arny collaboratively worked on the task until the participant receives a "Task is over" message on the screen after 13 collaboration cycles. This number of cycles was set by the research team with consideration of the recording time limit in Affectiva and the average amount of time participants spend on each cycle. However, the participants were not informed about the required number of the collaboration cycles before the task. This decision was to reduce the potential bias awareness of the study length could have on the participant's engagement. After task one was completed, the participant received the description for the second task. Similar to task one, the participant and Arny collaboratively worked on this task until the participant receives a "Task is over" message on the screen after 13 collaboration cycles. Affectiva emotion reports were monitored during both tasks in real-time and recorded for future analysis. Additionally, every collaboration task was screen recorded, and some notes and comments were taken by the wizards. Each collaborative design produced was also recorded for later evaluation.

After the two tasks, an interview and retrospective self-report process similar to Arny V1 study was followed to collect additional information about the participant's experience in each cycle of both controlled and treatment conditions. Audio recording of the retrospective protocols and interviews were captured for future analysis. Finally each task experience was rated by the participant using a 1-5 scale Likert scale.

6.1.4 Data Collection

Eight categories of data were collected during Arny V2 evaluation study:

1. Affectiva emotion reports for both tasks
2. Screen recording of the collaborative tasks
3. Wizards notes during the treatment and control conditions
4. Participants' self report during the mini-retrospective protocols for both tasks, and
5. Participants' response to the interview questions for both tasks
6. Collaborative interior design at the end of each task
7. Participants' Likert scale ranking of the controlled and treatment condition collaboration experience
8. Video and audio recording of an expert interior designer's feedback about the final designs.

The screen recording data from the treatment condition, Affectiva emotion reports, wizards' notes during the treatment condition, and the responses to interview question number one for the treatment condition provided the basis for responding to RQ1. Likert scale ranking of the two collaboration conditions, comparison of participants' responses to interview questions 2 to 4 and the retrospective protocol data from the two tasks were used to examine RQ2. Finally, the expert evaluation of the interior designs was used to explore RQ3. Throughout the following subsections, each data collection category is described in more detail.

1. ***Affectiva emotion records*:** Despite only using Affectiva emotion reports for the treatment task decision makings (Arny condition), Affectiva emotion reports were

captured during both tasks. The reports were used in real-time to direct the collaborative actions of the wizards during the treatment task. Also, the aligned screen and emotion recordings of both treatment and controlled condition were used post-study for analysis purposes, including for triangulation with the interview and retrospective data.

Similar to the exploratory study, although only specific affect dimensions were used for interaction with the participants, additional emotion dimensions were recorded along with valence and engagement values.

2. ***Screen recording of the collaboration:*** Affectiva's screen recording capability was used to record the collaboration process of both the control and treatment conditions. Similar to Arny V1 study, during the retrospective protocol the recordings were speed played to the participant to help them recall their actions and report their feelings during each cycle of the two collaborative tasks. The same screen recordings were also used for the RQ1 analysis. Using screen recordings along with mini-retrospective protocol responses, allowed this research to examine the participants' perception of Arny's actions in order to explore RQ2.

3. ***Wizard's notes about the participants actions and emotions:*** EIW tracked wizards' perception of the participant's actions type, Arny's action type, the contributed objects, and the participants valence and engagement level during both treatment and controlled conditions. To capture these details EIW utilized a revised version of the table used in Arny V1 study. In addition, the wizard added side notes

to the table of participant comments or questions that they shared through the microphone. Table 6 shows the structure followed by wizards for note taking.

	Cycle n		*Cycle n+1*	
	Participant	*Arny*	*Participant*	*Arny*
Contributed Object				
Wizard's Perception of Action Type: Diverged, Converged or Passed				
Wizard's perception of Affectiva Valence Category (Pos, neg, neut)				
Wizard's perception of Affectiva engagement report (Medium-high or low)				

Table 6- Structure followed by wizards for taking notes during Arny V2 study

4. ***Mini-retrospective reports:*** Following completion of tasks the 2.5x speeded recordings of the collaboration cycles were played to the participants, and the participant was given time to narrate each cycle. The interview questions were repeated to the participants after replaying each cycle. Participants' responses during this combined retrospective and interviews were 15 minutes on average for each condition. These self-reports were audio recorded and transcribed for the analysis. In order to answer the research questions, the retrospective data were analyzed and triangulated with Affective data.

5. ***Interview responses:*** The four interview questions were repeated to the participants after their retrospective report of each cycle of the collaboration. Participants'

answers were audio recorded and transcribed. Affectiva emotion reports were triangulated with respondents' answers to questions 1 from the treatment condition interviews. The result of the triangulation analysis was used to examine RQ1. Participants' responses to interview questions 2-4 for both conditions were used to study their perception of the collaboration for RQ2.

6. ***Collaborative designs:*** A screenshot of each collaborative interior design was captured at the end of each collaboration. These final designs were later reviewed by an expert interior designer as an input for RQ3 analysis.

7. ***Likert scale ranking of the two collaboration conditions:*** At the end of the retrospective protocol for each task, the participant was asked to rank their collaboration experience with the system during that task on a scale of 1-5, one being very negative and 5 being very positive. Comparison of the rating of the two conditions was used for RQ3 analysis.

8. ***Expert's review of the collaborative designs:*** All the final designs were evaluated by an expert interior designer. The expert came up with a set of metrics for comparing the final designs. She then compared the designs based on the extracted metrics. The aggregate scores for comparison of the two conditions were calculated based on these metrics in order to study RQ3. Details of the expert review metric factors are discussed in the analysis section.

6.2 Study Data

The research team collected four sessions of Arny V2 evaluation study as well as expert feedback on the final interior designs created during each session. The results showed significant differences between the two conditions in terms of user preference, engagement, and performance. Each participant's experience and a summary of the findings will be discussed in this section.

6.2.1 Study Participants

Arny V2 study sessions had to be stopped after collecting data from 4 participants as an impact of COVID-19 on face-to-face interactions. The study could not be continued online as the environmental factor with potential impact on participants' emotional feedback had to be consistent for all participants. Four Caucasian CS undergraduate students were recruited as the research subjects. There were 3 males and 1 female among the participants. The female participant's data had to be discarded after the study session due to specific facial expression characteristics. While conducting collaborative tasks, EIW noticed unusual patterns in this participant's Affectiva data. The situation was further investigated by asking the participant about her feelings during the retrospective phase while also closely observing her facial expressions in real-time. The real-time observations, as well as the comparison of the EIW notes of Affectiva emotion records versus self-reported affects, revealed that this participant smiled and laughed in response to unpleasant situations. As the Affectiva facial expression analysis was unable to provide an accurate evaluation of this participant's emotions, Arny V2's model of interaction was not properly followed during her treatment condition and her data had to be rejected. Meanwhile this

experience emphasizes the need for a pre-test to identify special cases with less expressive faces or specific unexpected expressions even with a homogenous user demography.

6.2.2 Data Analysis

Unlike Arny V1 exploratory study which was mostly focused on RQ1 & RQ2, during the Arny version 2 evaluation study analysis, a greater focus was placed on RQ3 which compares user AI creative collaboration with and without emotional feedback. Thus, in addition to data analysis similar to Arny V1 study, and asking the participants to rank the collaboration experience of the two conditions, a comprehensive expert review of the final designs was conducted.

In order to assess the final interior designs, an expert interior designer was presented with the collaboration condition task descriptions as well as final designs collaboratively created during each task. Each final design was marked with the participant number and the label that indicated if the design was an output from the faculty office interior design task or the startup manager office design task. The expert was however not aware of the interaction differences between the two conditions.

After reviewing the task description and the final interior designs, the expert introduced 1) Functionality of the space, 2) Aesthetics, 3) Circulation 4) Distribution/Organization, as metrics to evaluate each design. The expert also defined a metric called 5) Interaction to assess participant's handling of Arny unexpected contributions,

1. ***Functionality of the space.*** Space functionality is ranked based on convenience and usefulness of the space for the task client. This metric mainly focuses on how

well the interior design elements meet the space function described by the task description, i.e, how suitable is the space for a faculty or a startup manager wants.

2. ***Aesthetics.*** Aesthetics is ranked based on usage of decorative objects and the visual harmony of the space from the expert's point of view.
3. ***Circulation.*** Circulation is a metric used to assess how well an interior space is layed out so that it accommodates the movements of the users.
4. ***Distribution/Organization.*** Organization metric evaluates the grouping and relevance of the interior design elements, i.e., if the relative placement of two objects makes sense or not.

After reviewing each final interior design regardless of the collaboration specifics, the expert was provided with a sequenced list of Arny and participant contributions. The expert used this sequence for their rating of the interaction metric.

5. ***Interaction.*** Interaction is a creativity related factor described by the evaluator expert to measure how well the participant has handled unexpected objects that has been contributed by Arny. In conditions when the participant managed to make sense of an unexpected Arny contribution in a reasonable way, they received a higher interaction score from the expert.

6.3 Findings

Findings of Arny V2 study provide a comparison of collaboration with Arny V2 interaction model versus a control condition from the point of view of the users as well as

an outsider expert and the researcher. According to these findings, the 3 study participants preferred their interaction with Arny over the control condition. Researcher's analysis of Affectiva and participants' self-report of their feelings also confirmed that the users experienced more positive affect throughout Arny's condition. Finally, evaluations of the final interior designs by the third-party expert rated the designs produced during Arny condition as higher quality compared to the control condition final designs. The following subsections, provide a more detailed overview of these findings as well as a comprehensive report of each participant's collaboration experience.

6.3.1 Summarized Comparison of the Conditions

Findings of the Arny V2 evaluation study supported the design decisions that were recommended after the Arny V1 study. According to the 3 study sessions, Arny V2's interaction model based on valence and engagement was effective in maintaining a positive balance throughout the collaboration. In addition, similar to Arny V1 study, participants had a positive perception of the collaboration experience during the treatment condition compared to what they experienced in absence of emotional feedback. The evaluation study confirmed that the participants' positive responses were not correlated with specific action types- converging, diverging or pass. Instead, interview responses suggested that positive or negative emotions were tied with how well the participants' expectations from Arny were understood and met. These met or unmet expectations then triggered further judgments of each action and the overall collaboration in the participants:

Treatment Condition

Users rated both the collaboration with different conditions using a 5-point rating scale and all 3 participants preferred the collaboration with Arny (Treatment Condition). Facial expression data demonstrated high user engagement and positive valence compared to the control condition. Moreover, participants' description of their positive emotions was in line with Arny V1 results that there is a correlation between met expectations and positive valence.

Arny is Co-operative: The participants referred to Arny as a co-operative partner in the condition with emotion capturing. Results suggested that users perceived Arny's contributions related and complementary to their own contributions. They also said Arny influenced their contributions positively. Participants added that they felt associated with Arny in terms of their contributions. Arny created a story with them and communicated with the participants through its contributions. For example, in a scenario when the system converged to the participant's drawing of a window with drawing a frame, the participant stated:

> "I think it was interesting that the system basically did what I did. I just felt that it was interesting that it took something that I drew and did something else with it, that it responded accordingly."

Arny helps in design ideation with varied ideas: The result demonstrates that Arny helped the users in design ideation. Facial expression data suggested high engagement and positive valence even when Arny contributed a divergent idea. One of the participants stated that Arny's varying and divergent ideas helped them when they ran out of ideas. Even when Arny produced diverging ideas, participants perceived it as they were

still on the same task. Participants described some diverging action that they liked by statements such as:

> "The system was diverging...Sometimes I didn't have much idea on what to draw, but the system encouraged me to draw some next things. Here the system added a monitor and I could keep continuing by adding a keyboard and a mouse"

Control Condition

Users rated the control condition as least preferred in their ratings. The participants reported that Arny was contributing random objects to the office scene, which were "irrelevant" and "inappropriate". For example, Participant 2 was disappointed when he saw the system was drawing a dog in the office room. Two participants reported feeling so disappointed by the objects produced by the system that eagerness for achieving a good final design disappeared. Participants also mentioned the lack of influence and inspiration from the system. To describe the negative feelings participants used phrases such as "I didn't like it" or "it was annoying" and then described an unmet expectation. Then unmet expectations in case of converging contributions were described as a lack of creativity in the system, not understanding what the participant wanted, contributing an idea the participant was already done with, or leaving the participant helpless. Unmet expectations in reaction to diverging system contributions referred to the system as "random" and incapable of understanding what the user was doing.

In terms of the experienced affect, there was a higher occurrence of negative emotions across the control condition. In the few cases when the participants' experienced a positive emotion in this condition, the interview data demonstrated that it resulted from

the complementary and related contributions from Arny, or just being less challenging to the collaboration.

"There were times when the (control) system passed, and I felt so good, because I didn't have to figure out how to make it's drawings work."

For some participants, negative emotion led to interesting results. Participants 2 and 3 had initial negative emotions resulted from random inappropriate contributions of Arny. However, the initial negative emotion was followed by a sense of humor resulting in a creative solution to the unexpected situation. Further details will be provided in sections dedicated to session overviews.

6.3.2 Review of Individual Participant's Collaboration Experience

Review of each participant's experience will cover 3 primary areas: a) participants' feedback and perceptions of collaboration conditions, b) experienced affect during the collaboration, c) Analysis of their collaboration behavior and expert's review of their final designs.

Review of Participant 1 Collaborations

Participant ID: P1

Demographic details: Male undergraduate student in CS, Caucasian, 18-25 year age range

Self-reported level of design skills: 4

Order of tasks: Treatment condition (Arny V2) followed by control condition

Participant's Likert scale rating of each collaboration: Treatment condition 4 out of 5, Control condition 2 out of 5

a. Participants' feedback and perceptions of collaboration conditions

P1 summed up the system in treatment condition as a really helpful collaborator and capable of associating with his ideas. In contrast, he described the control system as not making much sense. The following is an example quote from P1 regarding his collaboration experience with Arny:

"... Like I would think: let's have a keyboard for that! And it was like: here you are… It is pretty cool that it is able to associate."

During the interview, when P1 was asked about Impact of Arny's contributions on his drawings, he indicated that he was impacted by Arny's contributions on multiple turns. For example, he noted about Arny's contribution on their second collaboration cycle

"It took out wall space (i.e. contributed an object suitable for wall space) so I made a few other things that took up wall space."

On a later diverging contribution, he commented:

"Yes (it impacted my drawing) because it moved from (objects for) one space which was fully done, to a new space, so I followed, I was now filling up the shelves."

Participant also shared positive feedback about impact of converging contributions from Arny on him. For example, on a turn when the participant contributed a monitor and Arny followed by drawing a keyboard he shared:

"Converging...I felt it went pretty well because we paired pretty well. It felt pretty good, ...(it impacted my later drawings) as from there I relied more on it (Arny)."

In contrast to the Arny condition, P1 described the control condition experience as:

"Some of the objects (it contributed) didn't make a lot of sense... they were irrelevant"

When P1 was asked about his overall score for the two systems, he scored the Arny condition as 4 and the control condition as 2 out of 5. Then he explained the only reason he is not rating the Arny as 5 is the drawing interface (i.e., Microsoft PowerPoint)

b. *Experienced affect during the collaboration*

P1's *interview responses reported more occurrence of both negative and neutral feelings during the control condition compared to the treatment condition.*

According to the interview self reports, P1 reported 8 turns with positive affect and only 2 turns with negative affect when collaborating with Arny when he reported only 3 turns with positive feelings but 6 turns with negative feelings when interacting with the control system (Figure 25).

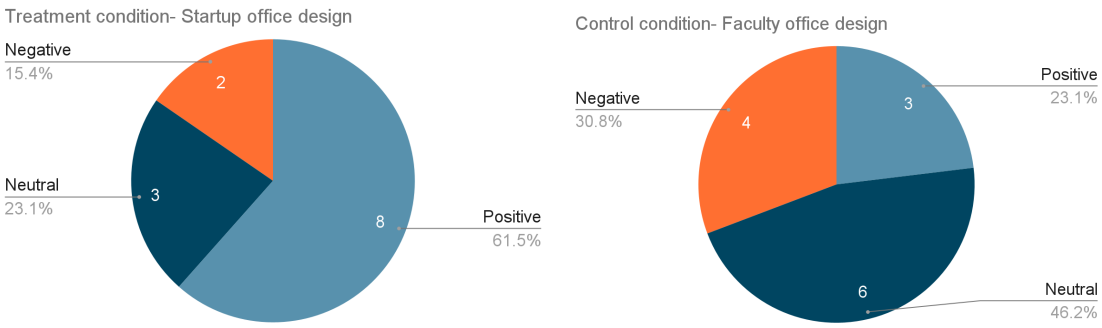


Figure 25- Experienced feelings according to P1 interview responses

An Affectiva engagement comparison of the Arny condition versus the control condition revealed the participant experienced higher engagement during the Arny

condition than during the control condition. According to these engagement reports, the participant experienced high levels of engagement for 10 out of the 13 turns during the Arny condition but only experienced high engagement for *2* out of the 13 turns of the control condition (Figure 26).

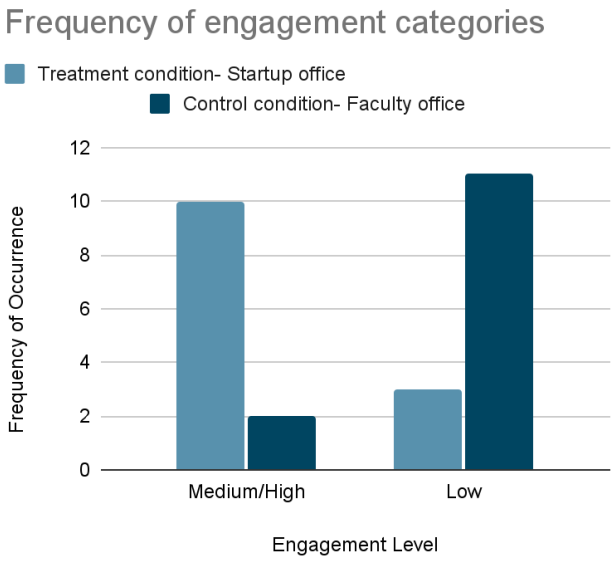


Figure 26- P1 engagement level during the two collaboration conditions based on Affectiva reports

c. Final interior designs

P1's collaborative interior design for the Arny condition received higher expert review scores when compared to the control condition. This participant received an overall score of 2.4 for his design in the Arny condition versus 1.4 for his design in the treatment condition. The breakdowns of the P1 scores for each condition are compared in Table 7 and Figure 27.

Condition	Overall Score	Function	Aesthetics	Organization	Circulation	Interaction

Treatment (Arny)	2.4	2	3	2	3	2
Control (Random)	1.4	2	2	1	1	1

Table 7- Expert review scores for P1 final interior designs from Arny and control conditions

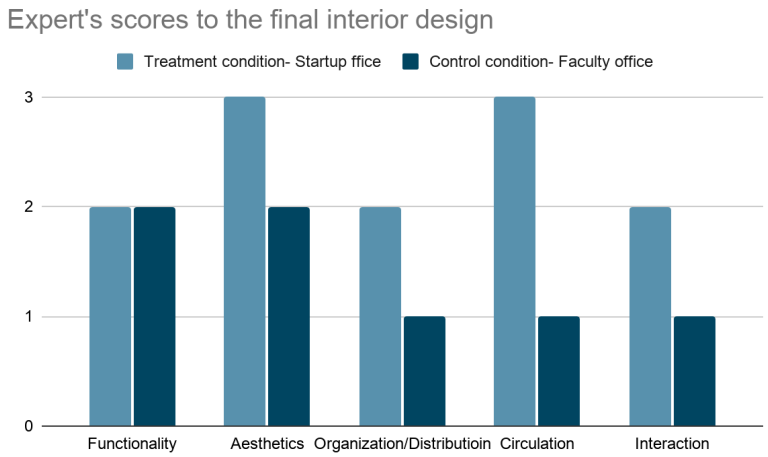
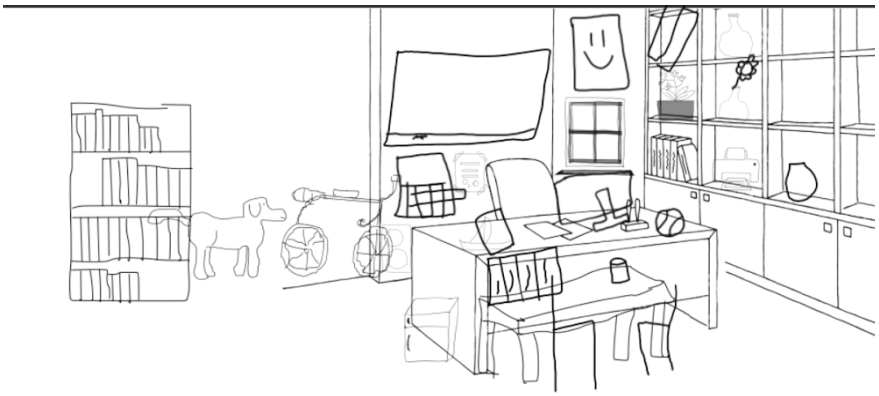


Figure 27- Evaluation of P1 final interior designs based on expert review

Figure 28 & 29 present the final interior designs created during the two collaborative tasks. An interesting observation when comparing the results is how P1 manages to sort out and place unexpected contributions from Arny during the treatment condition compared to the control condition. For example, when the participant was provided with a cut flower during the treatment condition, he followed by creating a vase for it and using it as a decorative object. The same participant however did not repeat a similar approach during the control condition and was frustrated in reaction to a decorative boat drawing that was contributed by Arny.

I added: socket

Figure 28- P1's final design for the Startup Manager office (Treatment condition)



I added: window

Figure 29- P1's final design for the Faculty Office (Control condition)

Review of Participant 2 Collaborations

Participant ID: P2

Demographic details: Male undergraduate student in CS, Caucasian, 18-25 year age range

Self-reported level of design skills: 4

Order of tasks: Treatment condition followed by control condition

Participant's Likert scale rating of each collaboration: Treatment condition 3 out of 5, Control condition 1 out of 5

a. Participants' feedback and perceptions of collaboration conditions

In describing the treatment condition, P2 referred to Arny as *"on the task"* and *"I understand why it did what it did",* whereas he described the control system as *"irrelevant."*.

During the interview phase, P2 noted that he found Arny cooperative throughout the process. When the participant described diverging actions by Arny, his explanation was that the system was "diverging but on task". The participant noted that the negative part of that experience was that visual perspective for the objects provided by Arny were limiting so he had to work with objects that couldn't be oriented as he liked. The participant explained that he overcame the challenge by resizing some of the "difficult to fit" objects and using them as decorative pieces on the shelves.

"It gave me a sofa, it made sense but the perspective wasn't right, so I made it tiny and put it on the shelf...It gave me another sofa and I put it on the shelf again"

Unlike the Arny condition, P2 described negative feelings and frustrations caused by diverging contributions from the control system. In spite of the control condition occurring after the treatment condition, the participant did not handle the unliked or difficult objects provided by the system with similar strategy. Instead of resizing difficult objects and placing them as decorative objects on the shelf area, the participant got confused and started distancing from the task goal. When the participant was asked about his experience during that collaboration, he replied:

"It gave me a dog; it didn't make any scene in a faculty office. I felt that it was not going to work anyways so I just added a bone….I couldn't do anything about it so I added the clock and the axe for fun."

b. *Experienced affect during the collaboration*

During the interview, P2's reported more occurrence of negative feelings across the control condition compared to the treatment condition. According to the interview self-reports, P2 reported 8 positive affective turns and 4 negative affective turns when collaborating with Arny, but only 2 positive affective turns and 8 negative ones when interacting with the control system (Figure 30). One interesting point about the way P2 described the positive and Neutral affect turns in the control condition was that when he lost trust in the system, he fabricated his own medieval story to distract himself from the negative feelings and have a better collaboration experience. P2 followed that this strategy also helped him stay engaged to the end of the task.

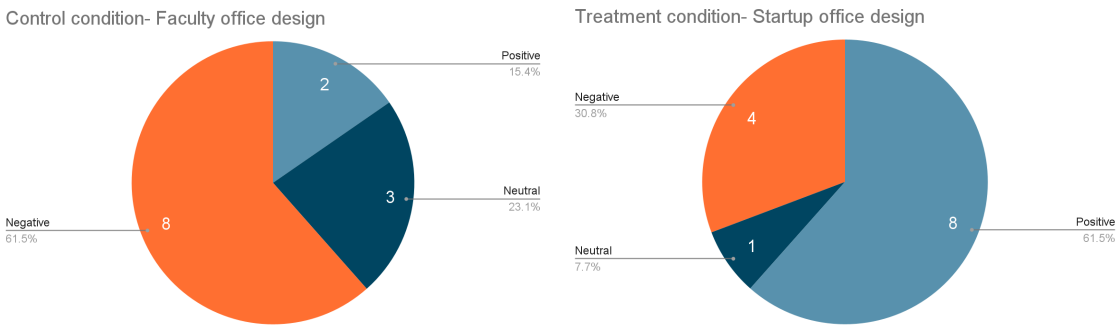


Figure 30- Experienced feelings according to P2 interview responses

Affectiva engagement comparison of the Arny condition versus the control condition for P2 is presented in Figure 31. P2's self-report explains how the higher level

of engagement in the control condition is not due to higher interest in that collaboration condition but resulted from his storytelling strategy to improve the task experience for himself.

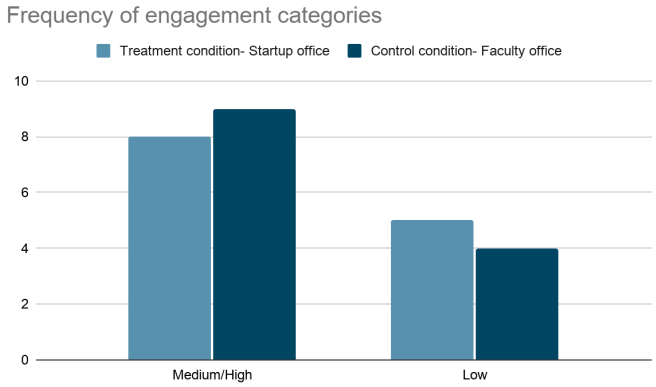


Figure 31- P2 engagement level during the two collaboration conditions based on Affectiva reports

c. Final interior designs

Expert review scores of P2's collaborative interior design were much higher for the Arny condition compared to the control condition. An overall score of 3.6 was given to P2 for his design in the Arny condition versus 1.4 for his design in the treatment condition. Table 8 and Figure 32 presents the breakdown of the P2 scores by condition.

Condition	Overall Score	Function	Aesthetics	Organization	Circulation	Interaction
Treatment (Arny)	3.6	3	4	4	3	4
Control (Random)	1.4	1	2	1	1	2

Table 8Expert review scores for P2 final interior designs from Arny and control conditions

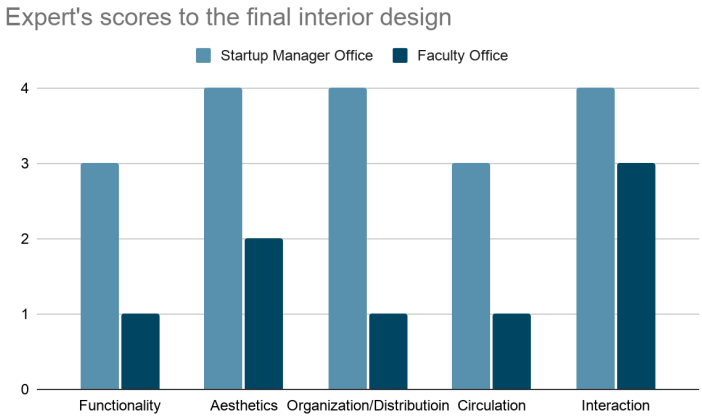


Figure 32- Evaluation of P2 final interior designs based on expert review

As discussed earlier about P2, this participant showed a significant change in interaction and strategy between the two conditions, which led to him creating a reasonable interior design in the Arny condition but losing interest in the task in the control condition. The difference between the two interior design outcomes can be viewed in Figure 33 & 34.

Figure 33- P2's final design for the Startup Manager Office (Treatment condition)

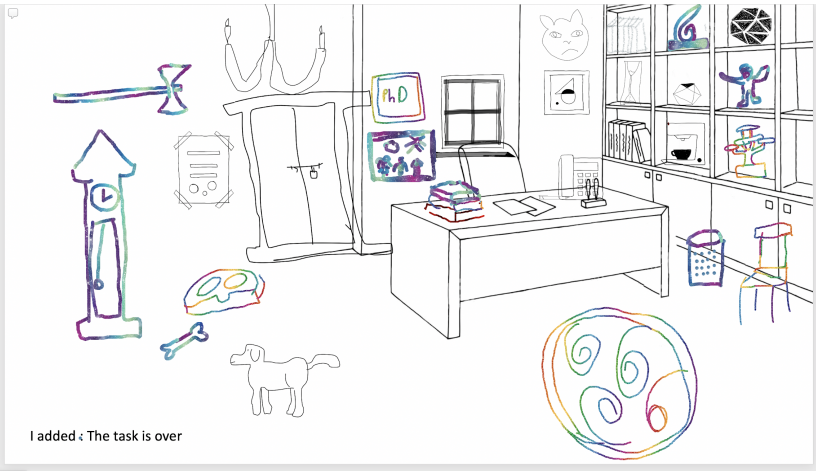


Figure 34- P2's final design for the Faculty Office (Control condition)

Review of Participant 3 Collaborations

Participant ID: P3

Demographic details: Male undergraduate student in CS, Caucasian, 18-25 year age range

Self-reported level of design skills: 5

Order of tasks: Control condition followed by treatment condition

Participant's Likert scale rating of each collaboration: Treatment condition 4 out of 5, Control condition 3 out of 5

a. Participants' feedback and perceptions of collaboration conditions

P3 describes the actions from the control system as *"random"* compared to Arny being a smarter system that *"built things around his ideas"*. P3 was thinking aloud throughout the experiment, and the researcher was remotely observing and listening to his interactions with the two systems which led to additional insights about his strategy, expectation and interpretation of the actions. In addition to the collaboration on the screen, this participant was talking to the system as working. During the control condition

interactions, he asked the system repeatedly, *"give me something useful this time."* or *"don't give me something random again!"*. He also once commented that *" I see I have to do all the work!"*. Another interesting behavior observed from this participant was objecting to the unliked contributions from the control system and then talking to the system about his strategy in response to those contributions. For example in an instance when the control system contributed a dog as a diverging action, the participant reacted:

"Common! A dog?!....I will put it further on the back…"

During the Arny condition on the other hand, P3 shared comments such as "this is good", *and* "it is working with me!" or communicated very specific expectations from the system such as:

"Give me another chair"

P3 also had conversations with system on his expectations of diverging contributions for more ideas or pass actions for freedom with comments such as:

" What should I draw? Don't pass on me now"

"Passed! Good!"

During the *retrospective* self-report, P3 shared that from his point of view the first system was random but the second system (Arny) was smarter and built around his plans. Then he stated:

"There were times it (Arny) diverged to expand my space of ideas."

And added:

"There were times I wanted to focus on my plans and the system (Arny) passed and it felt so good because I could focus and not worry about making something else work."

b. Experienced affect during the collaboration

P3 reported more occurrence of negative feelings across the control condition compared to the treatment condition. According to the interview self-reports, he reported 9 positive affective turns and 2 negative affective turns when collaborating with Arny, but 8 negative ones when interacting with the control system. In several instances during the control condition, P3 managed his emotional response to an unpleasant contribution from the system by creating a funny mini story or joke that helped him transition to a positive valence. He later explained the same behavior when describing his interaction emotions in his retrospective protocols. Breakdown of number of turns with different valence categories is presented in Figure 35.

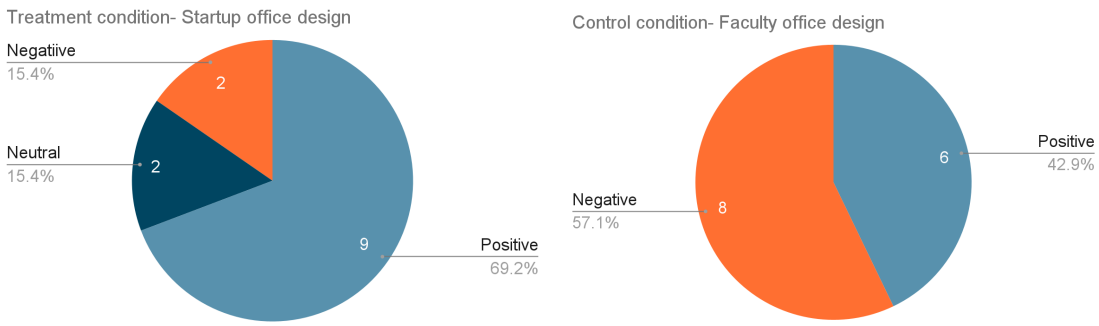


Figure 35- Experienced feelings according to P3 interview responses

Figure 36 presents the breakdown of the engagement level P3 experienced during each *collaboration* condition. According to the Affectiva reports the participant experienced much higher engagement during Arny condition with experiencing medium/High engagement for 10 turns during Arny condition but only 5 turns for the control condition.

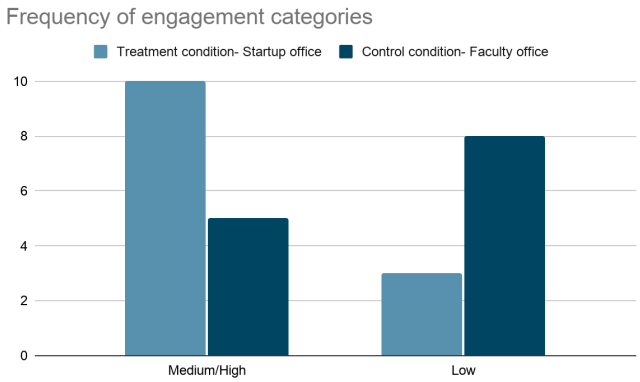


Figure 36- P3 engagement level during the two collaboration conditions based on Affectiva reports

c. Final interior designs

P3's collaborative interior design for the Arny condition received slightly higher expert review scores *when* compared to the control condition. This participant received an overall score of 3.8 for his design in the Arny condition versus 3.4 for his design in the treatment condition. The breakdown of the P3 scores for each condition are presented in Table 9 and Figure 37.

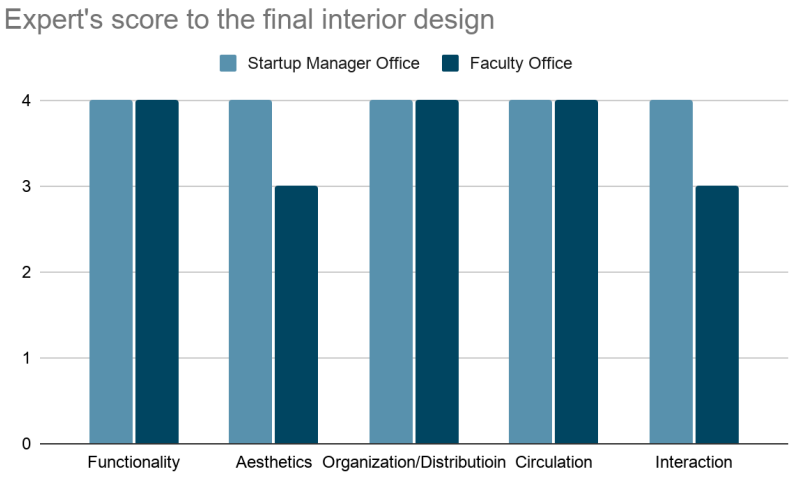


Figure 37- Evaluation of P3 final interior designs based on expert review

Condition	Overall Score	Function	Aesthetics	Organization	Circulation	Interaction
Treatment (Arny)	3.8	4	3	4	4	4
Control (Random)	3.4	4	3	4	4	3

Table 9- Expert review scores for P3 final interior designs from Arny and control conditions

In spite of P3's high expert review score in both conditions, an important aspect to note is the way this participant creatively dealt with some challenging contributions by Arny during the treatment condition. Transforming Arny's contribution of the rail and train to a projection on a wall monitor is an example of such creative solutions. Another notable contribution from this participant during the Arny condition was his glass wall idea to resolve the perspective issue with the sofa contributed by Arny. P3's final interior designs for both conditions are presented in Figure 38 & 39.

Figure 38- P3's final design for the Startup Office (Treatment condition)

Figure 39- P3's final design for the Faculty Office (Control condition)

6.4 Summarized Review of Arny V2 Evaluation Study

In this Chapter, an evaluation study of Arny V2 interaction model for co-creative collaboration was discussed. The study reviewed the participants' interaction experience with Arny versus a control system and compared their experience in terms of satisfaction from the collaboration, experienced affect across each collaboration as well as the quality of the final interior designs generated at the end of each collaboration. According to the study findings, all study participants reported higher satisfaction of their collaboration with Arny V2 versus their collaboration with the control system that was not interacting based on emotional feedback. Also, all participants experienced more positive valence values during the Arny condition versus the control condition. Finally expert evaluations rated the final interior designs generated in collaboration with Arny with a higher score compared to the control condition outcomes.

Chapter 7: CONCLUSION

Similar to human-human co-creative collaboration, human-AI creative collaboration can benefit from a negotiation mechanism that leads the collaboration dynamics as well as contributions. With this basis the following statement was made in the beginning of this dissertation:

Using knowledge of human emotion in human-AI co-creative collaboration can improve the user satisfaction of the collaboration experience and quality of *the collaboration outcome.*

To validate this statement, this dissertation proposed an interaction model for co-creative collaboration based on emotional feedback and explored the following three research questions around that model:

RQ1: What are the design parameters of an emotion-based interaction model for co-creativity?

RQ2: What is the human perception of co-creativity with emotion detection?

RQ3: How are outputs different between interactions with and without emotion detection?

In this dissertation document, that interaction model is described, evaluated, and revised in response to these research questions.

7.1 Arny Model of Co-creativity Collaboration

To explore the thesis statement through the research questions, Arny, an interaction model incorporating emotion detection in co-creative design with AI was presented and two versions of this model were deployed. Through the iterative design of Arny, RQ1 was answered and the following design parameters were defined for an emotion-based interaction model for co-creativity:

1. Emotion stimuli: Arny V2 targets the external emotions triggered by user's judgment of the collaborator actions or future expectations from the collaborator
2. Capturing user's affect: Arny's model of interaction relies on Affectiva emotion recognition based on facial expression to passively capture user affect without distracting the flow of the collaboration
3. Affect interpretation: Arny V2, relies on valence and engagement and provides a structure for how to target meaningful periods of user's emotional feedback for the purpose of the collaboration.
4. Interaction model: Arny V2 interaction model includes a collaboration model that specifies how to separate the emotion stimulus elicited by the collaborator and a set of collaboration rules on how to respond to this emotional feedback.

Both versions of Arny's interaction models follow a turn taking pattern where Arny and user each make a single contribution on their turns. The first version, Arny V1, relied on the collaborator's valence level in response to Arny's last contribution for selecting a convergent or divergent component to contribute to the design or to pass on the AI

contribution. This approach allowed Arny V1 to interact based on a judgment of the user's stance in response to its latest contribution in the collaboration. The evaluation study of Arny V1 suggested consideration of a second affect parameter, engagement, in order to predict the user's future expectation from the AI in addition to understanding their stance towards the past AI contribution. The rules that govern the AI contribution in Arny V2 take into account engagement level as well as valence response during the last cycle of collaboration.

Arny V1 and Arny V2 studies also explored RQ2: What is the human perception of co-creativity with emotion detection? Exploratory study of Arny V1 showed that users were generally happy with collaborating with an AI partner and they experienced a mostly positive emotional experience [Abdellahi et al, 2020]. In addition, the study concluded that it is important to amend the interaction model and to include an engagement factor to the interaction model of Arny V2.

Arny V1 exploratory study was followed by a evaluation study using Arny V2 that had a baseline (control) condition to compare with users' interaction experience with Arny. During the control condition the AI partner randomly chose whether to make converging or diverging contributions, while during Arny condition the AI partner applied Arny V2 interaction rules based on emotional feedback instead of a random selection. The participants were more satisfied in the Arny condition and expressed in the debriefing that they felt that the AI partner in Arny condition was responsive to their needs and expectations as a creative collaborator.

Finally, to explore RQ3 on outputs different between interactions with and without emotion detection, the collaboration outcomes from the two conditions of Arny V2 study

were evaluated and rated by an expert designer. Comparison of the ratings for the design outcomes of Arny condition versus Control condition demonstrated a higher quality of outcome when a model of interaction based on emotional feedback was followed.

7.2 Emotion and Engagement Detection Technology Limitations

A number of restrictions were imposed by the technologies available for the research and design of Arny V1 and V2. The primary challenges for both studies were due to limitations of the emotion recognition technology. Currently available emotion recognition technologies provide less accuracy for non-Caucasian users due to baseline databases that are predominantly Caucasians. Additionally, less expressive users present challenges for reading emotions. Such underlying factors limit the effectiveness of interaction models based on emotional feedback due to inaccurate emotion recognition. Also, the 0 valence value reported by the Affectiva application when it could not read the user's face due to their movement outside the frame contributed to the shortcomings of the Arny studies.

A further limitation encountered during the Arny V2 case studies was the eye tracking technology that was initially included in the study setup for additional data collection on users' attention focus. The eye tracking technology available to this research could not capture the full range of motion for users, thus we discarded that plan for Arny v2 study. By using more advanced eye tracking technology, Arny's interaction model can be enhanced to consider a history of contributions rather than simply targeting the latest ideas for converging or diverging actions.

7.3 Data Collection Limitations

In addition to the technology limitations, this dissertation research has been limited by the challenges faced during the public health emergency caused by COVID-19. Due to the safety and welfare of the human participants and the research staff involved in this dissertation research, and according to the instructions provided by UNC Charlotte IRB, this dissertation's case study using Arny V2 had to be paused starting March 2020. In light of the fact that emotional responses are largely influenced by external contexts, it was not possible to continue the case study research remotely during pandemic expansion. As a consequence, the data provided in this report only describes the observations for a limited number of participants. Consequently, it is possible to observe more variation of emotional and behavioral responses in a co-creative design activity using Arny V2 with more participants.

7.4 Future Research

More research is needed to establish a comprehensive interaction model and system design for human AI collaboration as well as to understand user's perceptions and expectations in the context of such collaboration. To provide an improved model based on Arny's interaction model, future work should explore the user's concentration focus using technologies such as eye tracking glasses with motion freedom rather than relying on the assumption user affect is elicited by the recent collaborator contributions. Similarly, to capture the user's affect more accurately, it is vital to use technology that has less constraints on users' range and space of motion and provides accurate emotion reporting for different races, ethnicities and expressiveness levels. An alternative to address the

expressiveness impact on emotion reading is for the emotion reporting to be calibrated based on a pre-task.

To progress the research presented in this dissertation, future research can also examine the impact of AI's action sequences on the user in addition to the interaction model presented in this research. Future post-pandemic research must examine a broader range of case studies for themes related to user behavior, expectation, and perception in addition to revising the interaction model and technology.

REFERENCES

1. Abdellahi, S., Rezwana, J., Algarni, A., Maher, M. L. (2019). The Co-Creation Behavior Framework: An Empirical Investigation of Collaborative Drawing. Heron Island Conference Computational and Cognitive Models of Creative Design.
2. Abdellahi, S., Maher, M. L., Siddiqui, S., Rezwana, J., & Almadan, A. (2020, July). Arny: A Study of a Co-creative Interaction Model Focused on Emotion Feedback. In *International Conference on Human-Computer Interaction* (pp. 377-396). Springer, Cham.
3. Abdellahi, S., Maher, M. L., & Siddiqui, S. (2020). Arny: A Co-Creative System Design based on Emotional Feedback. In *ICCC* (pp. 81-84).
4. Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. Review of general psychology, 10(3), 229-240.
5. Clandinin, D. J., & Connelly, F. M. (2004). *Narrative inquiry: Experience and story in qualitative research*. John Wiley & Sons.
6. Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J., & Cottrell, G. W. (2010). Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion*, *10*(6), 874.
7. Darwin, C., & Prodger, P. (1998). The expression of the emotions in man and animals. Oxford University Press, USA.
8. Davis, N. M., Hsiao, C. P., Singh, K. Y., Li, L., Moningi, S., & Magerko, B. (2015, June). Drawing Apprentice: An Enactive Co-Creative Agent for Artistic Collaboration. In Creativity & Cognition (pp. 185-186).

9. De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. Phenomenology and the cognitive sciences, 6(4), 485-507.
10. DiPaola, S., & McCaig, G. (2016, July). Using Artificial Intelligence Techniques to Emulate the Creativity of a Portrait Painter. In EVA.
11. Doerrfeld, B. 2015. https://nordicapis.com/20-emotion-recognition-apis-that-will-leave-you-impressed-and-concerned/. Retrieved 25 April 2019.
12. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects?. *ACM Transactions on graphics (TOG)* 31(4), 1-10 (2012).
13. Ekman, P. (1999). Basic emotions. Handbook of cognition and emotion, 45-60.
14. Eligio, U. X., Ainsworth, S. E., & Crook, C. K. (2012). Emotion understanding and performance during computer-supported collaboration. Computers in Human Behavior, 28(6), 2046-2054.
15. Fuller, D., & Magerko, B. (2010, June). Shared mental models in improvisational performance. In *Proceedings of the intelligent narrative technologies III workshop* (p. 15). ACM.
16. Fuller, D., & Magerko, B. (2011, November). Shared mental models in improvisational theatre. In *Proceedings of the 8th ACM conference on Creativity and cognition* (pp. 269-278). ACM.
17. Glines, P., Griffith, I., & Bodily, P. M. (2021) Software Design Patterns of Computational Creativity: A Systematic Mapping Study.
18. Gross, J. J., & Levenson, R. W. (1993). Emotional suppression: physiology, self-report, and expressive behavior. Journal of personality and social psychology, 64(6), 970.

19. Gumienny, R., Gericke, L., Wenzel, M., & Meinel, C. (2013, February). Supporting creative collaboration in globally distributed companies. In Proceedings of the 2013 conference on Computer supported cooperative work (pp. 995-1007). ACM.
20. Holstein, J. A., & Gubrium, J. F. (Eds.). (2011). *Varieties of narrative analysis*. Sage Publications.
21. Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. International journal of human-computer studies, 59(1-2), 1-32.
22. Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. Proceedings of the National Academy of Sciences, 109(19), 7241-7244.
23. Jacob, M., Zook, A., & Magerko, B. (2013). Viewpoints AI: Procedurally Representing and Reasoning about Gestures. In DiGRA conference.
24. Jordanous, A. (10 April 2014). "What is Computational Creativity?". https://www.creativitypost.com/index.php?p=science/what_is_computational_creativity, Retrieved 23 April 2019.
25. Kapoor, A., & Picard, R. W. (2001, November). A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces* (pp. 1-5). ACM.
26. Karimi, P., Maher, M. L., Grace, K., & Davis, N. (2018). A computational model for visual conceptual blends. *IBM Journal of Research and Development*.
27. Kellas, J. K., & Trees, A. R. (2005). Rating interactional sense-making in the process of joint storytelling. *The sourcebook of nonverbal measures: Going beyond words*, 281.

28. Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and emotion, 5(4), 345-379.
29. Kohler, T., Fueller, J., Stieger, D., & Matzler, K. (2011). Avatar-based innovation: Consequences of the virtual co-creation experience. Computers in Human Behavior, 27(1), 160-168.
30. Magdin, M., & Prikler, F. (2018). Real Time Facial Expression Recognition Using Webcam and SDK Affectiva. *International Journal of Interactive Multimedia & Artificial Intelligence*, *5*(1).
31. Makhaeva, J., Frauenberger, C., & Spiel, K. (2016, August). Creating creative spaces for co-designing with autistic children: the concept of a Handlungsspielraum. In Proceedings of the 14th Participatory Design Conference: Full papers-Volume 1 (pp. 51-60). ACM.
32. Matthews, G., Davies, D. R., Stammers, R. B., & Westerman, S. J. (2000). Human performance: Cognition, stress, and individual differences. Psychology Press.
33. McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016, May). AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems* (pp. 3723-3726). ACM.
34. McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J., & Picard, R. (2013). Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 881-888).

35. Mikolov, T., Chen, K., Corrado, G, Dean, J.: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, (2013).
36. Miyake, Y. (2002). Co-creation system. Cognitive Processing, 3, 131-136.
37. Morgan, J. T., Gilbert, M., McDonald, D. W., & Zachry, M. (2014, February). Editing beyond articles: diversity & dynamics of teamwork in open collaborations. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 550-563). ACM.
38. Paulus, P. B., Levine, D. S., Brown, V., Minai, A. A., & Doboli, S. (2010). Modeling ideational creativity in groups: Connecting cognitive, neural, and computational approaches. Small Group Research, 41(6), 688-724.
39. Picard, R. W., & Daily, S. B. (2005, April). Evaluating affective interactions: Alternatives to asking what users feel. In *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches* (pp. 2119-2122). New York, NY: ACM.
40. Plutchik, R. (1991). *The emotions*. University Press of America.
41. Prinz, J. (2005). Are emotions feelings?. *Journal of consciousness studies*, *12*(8-9), 9-25.
42. Rehurek, R. and Sojka, P.: Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer, (2010).
43. Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.
44. Sacharin, V., Schlegel, K., & Scherer, K. R. (2012). Geneva emotion wheel rating study.

45. Sanders, E. B. N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. Co-design, 4(1), 5-18.
46. Sawyer, R. K. (2014). *Group creativity: Music, theater, collaboration*. Psychology Press.
47. Sawyer, R. K., & DeZutter, S. (2009). Distributed creativity: How collective creations emerge from collaboration. *Psychology of aesthetics, creativity, and the arts*, *3*(2), 81.
48. Scherer, K. R. (2005). What are emotions? And how can they be measured?. Social science information, 44(4), 695-729.
49. Shneiderman, B. (2007). Creativity support tools: Accelerating discovery and innovation. *Communications of the ACM*, *50*(12), 20-32.
50. Visscher, K., & Fisscher, O. A. (2009). Cycles and diamonds: how management consultants diverge and converge in organization design processes. Creativity and innovation management, 18(2), 121-131.
51. Xu, T., White, J., Kalkan, S., & Gunes, H. (2020, August). Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision* (pp. 506-523). Springer, Cham.
52. Rouse, M. "Computational Creativity". https://whatis.techtarget.com/definition/computational-creativity. Retrieved 23 April 2019
53. Affectiva Help Center, Retrieved March 2019.
54. Teamviewer Remote Access and Remote Control Application, https://www.teamviewer.com/en-us/

55. Ziteboard Online Whiteboard with Realtime Collaboration, https://ziteboard.com/

APPENDIX 1: CASE STUDY PRE-SCREENING QUESTIONNAIRE

Please respond to the following questionnaire to help us evaluate if you are eligible to participate in our study. You will receive a follow-up by the study team **ONLY** if we find you eligible.

Your email address will be recorded when you submit this form.

* Required

What is your age? *

- Under 18
- 18-35
- Over 35

What is your gender *

- Male
- Female
- Other

What is your level of education? *

- High school or lower
- Undergraduate student or Bachelor degree
- Master student or Masters degree
- PhD student or higher

Did you ever had any design or drawing training or education? *

- Yes
- No

On a scale of one to five, one being very low and five being very high, how comfortable you are with drawing? *

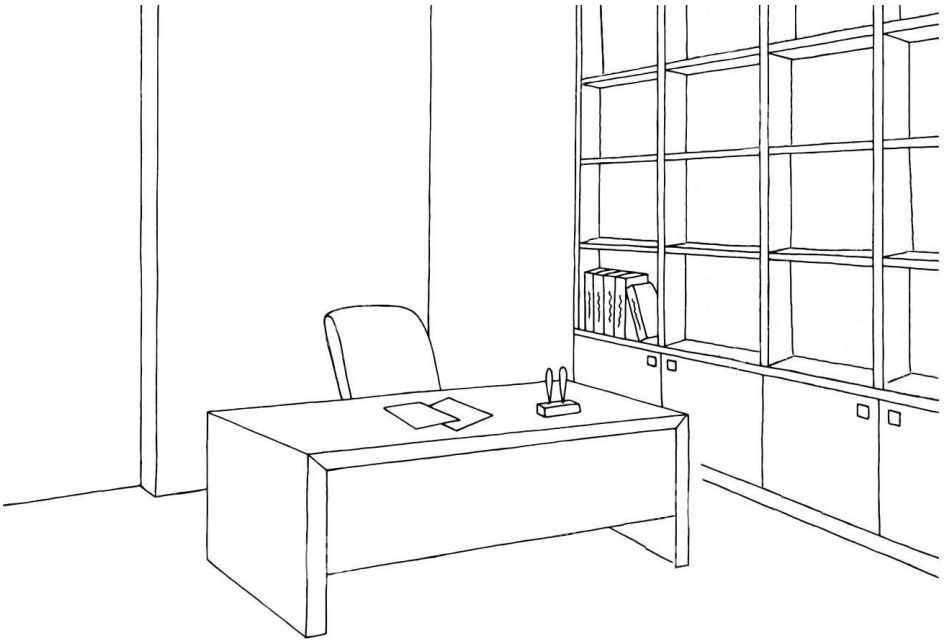
Very Low 1_____2_____3_____4_____5 Very High

APPENDIX 2: ARNY V1 EXPLORATORY STUDY TASK DESCRIPTION

You are going to engage with Arny our Artificial Intelligence system on a design task. Arny will take turns with you, and you will indicate you are ready for Arny to take a turn by Placing a checkmark on the turn taking table on the right side of the screen.

Design task:

You are hired to design the interior of an office for the Director of a new tech start up. She will be using the office to meet with potential investors and with her team of creative tech developers. You can sketch objects to be placed on the shelves, the floor, the walls, and the desk.



Me	
Arny	
Me	
Arny	
Me	
Arny	
Me	
Arny	
Me	
Arny	
Me	
Arny	
Me	
Arny	
Me	
Arny	
Me	
Arny	
Me	
Arny	

APPENDIX 3: CASE STUDY PRE-SCREENING QUESTIONNAIRE

Please respond to the following questionnaire to help us evaluate if you are eligible to participate in our study. You will receive a follow-up by the study team **ONLY** if we find you eligible.

Your email address will be recorded when you submit this form.

* Required

What is your age? *

- Under 18
- 18-35
- Over 35

What is your gender *

- Male
- Female
- Other

What is your race/ethnicity? *

- Caucasian American
- Caucasian Not American
- African American
- Hispanic or Latino
- Asian
- Middle Eastern
- Other

Do you think you have an expressive face? *

- Yes
- No
- I don't know

What is your level of education? *

- High school or lower
- Undergraduate student or Bachelor degree
- Master student or Masters degree
- PhD student or higher

Did you ever had any design or drawing training or education? *

- Yes
- No

On a scale of one to five, one being very low and five being very high, how comfortable you are with drawing? *

Very Low 1_____2_____3_____4_____5 Very High

APPENDIX 4: CASE STUDY TASK DESCRIPTION

You are going to engage with our co-creative AI system, Arny, on two interior design tasks. You are going to start the tasks by adding the first object to the design. Arny will take turns with you, and each of you will be adding one object to the interior design on your turn. To help Arny understand your contribution better, you will type the name of the last object you contributed to the office design at the end of your turn and Arny will follow the same pattern. To make it easier for Arny to understand your contribution, please try to use common one-word names for your objects when possible.

Please consider the follow points while interacting with Arny:

- Make sure to type the name of the object you contributed at the END of each turn and with a ONE-WORD label (except if more than one word is necessary)
- Do not move the mouse when it is not your turn
- Help Arny by resizing and relocating the objects contributed by the AI
- You must contribute ONE object on your turn. Arny however might "pass" a turn if it has trouble choosing an object.
- The task is focused on interior design and the type of objects you contribute and not the quality of your drawing.

In addition, please note that Arny is still in very early development stages. The version of Arny you see during the current study uses Arny's AI backend with help of Microsoft PowerPoint as a temporary front end. So, while interacting with the system, please only focus on how well Arny understands you and collaborates with you and ignore any challenges you might face with the interface or response speed.

Interior Design Task 1:

You are hired to design the interior of an office for the Director of a new tech start-up. She will be using the office to meet with potential investors and with her team of creative tech developers. You can sketch objects to be placed on the shelves, the floor, the walls, and the desk.

Interior Design Task 2:

You are hired to design the interior of an office for a university professor. She will be using the office to meet with her students and colleagues, prepper for her classes and work on her research projects. You can sketch objects to be placed on the shelves, the floor, the walls, and the desk.