An experimental multi-modal approach to instrument the sensemaking process at the team-level

by

Ignacio Salvador Vazquez Rodarte

M.Sc. Automotive Technology, TU Eindhoven (2015)
B.Sc. Mechatronics Engineering, Tec de Monterrey (2011)

Submitted to the System Design and Management
in partial fulfillment of the requirements for the degree of

Master of Science in System Design and Management

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author . . . .
System Design and Management
August 10, 2022

Certified by . . . .
Bryan Moser
Sr. Lecturer and Academic Director, System Design and Management
Thesis Supervisor

Accepted by . . . .
Joan Rubin
Executive Director, System Design and Management Program

An experimental multi-modal approach to instrument the sensemaking process at the team-level

by

Ignacio Salvador Vazquez Rodarte

Submitted to the System Design and Management
on August 10, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in System Design and Management

Abstract

Just like Julius Caesar's Gaul, any engineering challenge can be divided into three parts; (1) the problem, (2) solution and (3) design spaces. The interaction between solution and design, and the degree of influence that any given team has upon them will depend on the capacity of said team to make sense of the problem. This thesis presents a framework to evaluate the process of team-level sensemaking. How a small group of individuals show emotion, converse with each other and interact with the engineering problem at hand. This integrated view is tested with a small-n experiment to demonstrate the possible insights and data that can be generated and analyzed. As a contribution to collective intelligence and teamwork, the ability to objectively judge a team's performance —via Pareto Ranks, measure the conversation dynamics —using graph theory and voice recognition, assessing the average emotion content displayed by the team members —using facial recognition, and estimating team entanglement —with physiological signals captured by smartwatches, gives a deep dive into each team's sensemaking process.

Pending reproduction of the experiment, this first iteration seems to indicate that emotions play a role in a team's motivation to perform, as does the timing of the conversations the team has. Also, there are heuristics that emerged from the teams when they had to judge which of their proposed in-game designs was better —even though none of them actually met the requirements.

The work presented in this thesis is the enactment of one specific sensemaking framework: Weick's seven properties. Applied to a bounded system, where the problem under analysis is fully understood and the teams operate in game-bubble, where they all have access to the same —yet purposefully limited— information. And just like the participants in this experiment, this thesis might not have crossed the finish line set by the goals, but it certainly moved closer to it.

Thesis Supervisor: Bryan Moser
Title: Sr. Lecturer and Academic Director, System Design and Management

Acknowledgments

In philosophy, an aporia is a conundrum or state of puzzlement, while in rhetoric it is a declaration of doubt. That is exactly how this thesis got started and also the state at which I found myself in multiple times. Just like the stories of old, this journey could not have been what it became without a rather large group of people that knowingly or not made my time at MIT possible.

First and foremost, my parent and my sister, whom throughout the years have served me as a distant source of calm and have kept me grounded. And since 24-hours ago [1], my baby niece, Carolina. Welcome to the family.

My thesis-mates Jakob Kruse and Shio Peng, who worked just as hard during the last few months to create an experiment that could serve the three of us. I enjoyed working with you and I hope you found all of our discussions useful. This piece of work has been undoubtedly influenced —for the better— by your points of view.

Thesis-writing was not the only thing I did during the year at MIT. A special shout-out goes to the SDM TA Team; Yuliya, Bruce, Ricardo, Ben and Yuru. I can only wish to get at leat a *meets expectations with minor flaws* were you the ones grading this piece of work.

I am really glad to have had the chance to continue the work laid out by previous students on the Global Teamwork Lab; Lorena, PS, Kevin, Catherine, Prakash and Nate. Hats-off to you all. Also, Dipti, thank you for saving me during the summer getting the simulation model to work.

My time at MIT actually started long before I came to the US. Back in the Netherlands, I have to recall my TMC fellow emploeyeneurs, Bram, Antonio, Gijs, both Marias, Ruben and Romain. Thanks for your friendship.

Last but not least, my closest friends, who no matter how hard I've tried to get rid of you, you have prevented me for doing so. Francisco, Karime and Mariangel.

[1] As per the time I am actually writing this section.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> If you only do what you can do, you will never be more than what you are now.
>
> Master Oogway, *Kun-fu Panda* (2008).

Without giving away any spoilers, the most common starting point when people need to make sense of a complex situation is *confusion* [Weick, 1995]. And it is in the same spirit that this chapter aims to provide an overview of the research work presented in the thesis. Chapter 1 is the starting point to understand why furthering scientific understanding about sensemaking is relevant (**Section 1.1**).

A summary of the thesis content chapter by chapter is available in **Section 1.2**.

Without further ado, it is time to go find that white whale.

1.1 Research Motivation

Sensemaking, first introduced by Karl Weick [Weick, 1979], is currently used in fields ranging from leadership [Ancona et al., 2007], strategy [Kaplan & Orlikowski], organizational development [Kaplan & Orlikowski], pedagogical science [Stigliani & Ravasi, 2012] and product development [Wright et al., 2000]. While all the fields have adapted approaches to improve people's and organizations' capability to engage in sensemaking, there is a lack of research about methods to measure sensemaking in real-time and in an unobtrusive way adequately.

Definition 1 (Sensemaking)**.** The process by which people turn information and the circumstances of a situation explicitly into words and that serves as a springboard into action [Weick, 1995].

By establishing a method to measure sensemaking, new opportunities become available to detect patterns across disciplines [Maitlis & Christianson, 2014] and characterize intermediate and multi-level processes [Stigliani & Ravasi, 2012]. Moreover, a coherent prescriptive version of sensemaking across scales could be created by building a descriptive catalog of sensemaking events, practices, behavior, and contexts.

The need for said catalog is a common concern of engineering practitioners: to have an ex-ante set of rules and practices that help them guide their engagement with novel and —usually— ill-defined complex systems. Even though there is no lack of good practices proposed by organizational research regarding communication equity [Kim et al., 2008], balancing the personalities of a team based on individual personality features [Riedl et al., 2021], the influence of experience and specialization for a given task [Hindmarsh & Pilnick, 2007] and the use of positive and negative emotion in a team's performance [Maitlis et al., 2013], sensemaking is often abstracted as a concept and dissociated from the context at hand, in particular, separate from the nature of the problem being worked on and available solutions at the moment. In part, this gap in the research literature yields a challenge of properly instrumenting the way people work in today's distributed environment ([Weick et al., 2005], [Maitlis & Christianson, 2014], [Kozlowski & Chao, 2018]).

For this research, teams are the relevant and meaningful unit of analysis. There remains a need for a theory on collective sensemaking informed by understanding sensemaking as a shared emotional and distributed process [Grandey, 2008]. The instruments and signals associated with the sensemaking features must be able to work for groups of people and individuals [Stigliani & Ravasi, 2012].

Given that a well-accepted property of sensemaking is that it relies heavily on the use of language and reflection, much research has focused on (micro-) ethnographic studies ([Ancona, 2012], [Stigliani & Ravasi, 2012]) or embedded-observer approaches ([Maitlis, 2005], [Liu & Maitlis, 2014]). These methods have a shared strength of analyzing group's narratives and their patterns [Maitlis & Christianson, 2014]. They also face a significant challenge, as does the study presented in this thesis, of scaling up and allowing for a smooth transition from individual to organization and back.

The relationship between the composition of design teams and the problem being worked on has been identified as an influencing factor in how a team engages with a system [Fruehling & Moser, 2018]. This paper will refer to the previous two perspectives as *Social Space* and *Problem Space*.

Definition 2 (Social Space)**.** The mapping of how people relate to one another, their utility (of value in different dimensions) and their position and power relative to one another.

Definition 3 (Problem Space)**.** Includes elements that could be directly or indirectly affected by the proposed system solution or that, in turn, have an exogenous influence on the chosen solution. Said influences are often governed by a set of interconnected stakeholders' needs —expressed as goals and targets coupled with the fabric of the problem that provides constraints to the system.

Meanwhile, the final metrics used to evaluate the decisions of the team will be known as *Solution Space*.

Definition 4 (Solution Space)**.** The space of variables, from the system being acted upon, that the decision-making team can alter.

This research assumes that in order to have a complete picture of sensemaking, measurements across all three spaces are necessary [Pelegrin et al., 2018]. The fact that the *Solution Space* is uncertain and unknown is a significant reason sensemaking is necessary [Ancona, 2012]; thus, making the Spaces-triad a sociotechnical system.

Given the relevance of the intrinsic qualities of the Problem Space, a formal approach to defining and quantifying its level of complexity is also needed [Carroll, 2021]. Additionally, the interaction between the actors/agents and the environment is an integral part of the process by which knowledge is constructed [Weick, 1979] — by taking action and seeing the effect on the system's response. Finally, even though a

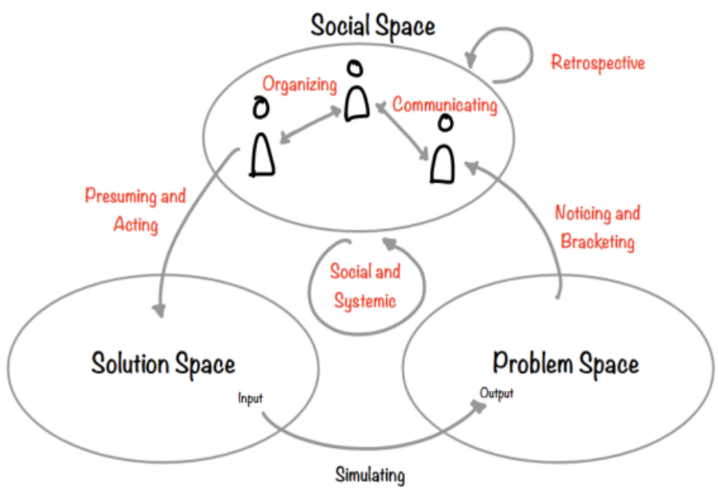


Figure 1-1: Design Spaces and Weick's Sensemaking Properties diagram. This is a conceptual representation of the relationship between the Social, Problem and Solution Spaces; and the sensemaking descriptive properties [Vazquez et al., 2022].

great deal of literature focuses on the theory of sensemaking [Maitlis & Christianson, 2014], there are concrete ideas to instrument processes that indicate sensemaking is being carried out, such as conversational pattern identification [Basu et al., 2001], labeling and categorization of new information [Wolbers, 2021], use of physical artifacts ([Stigliani & Ravasi, 2012], [Yang, 2005]) and generation of hypotheses [de Weck et al., 2011].

The development of methods and instruments to characterize the structure of organizations has also been used to find innovative teams [Gloor et al., 2008]. However, sensemaking is a process through which people work to understand issues or events that are novel, ambiguous, confusing, or in some other way violate expectations, and by its very nature, it is ephemeral. Measuring the process has remained a challenge.

1.2 Thesis Outline

Chapter 1 introduces the basic context about sensemaking, the relevance of this research and the research questions addressed.

Chapter 2 presents the previous research done in the fields used during this research; (1) sensemaking and (2) teamwork.

Chapter 3 links the literature background to specific research objectives and hypotheses. It also lays the fundamental concepts that guide the experiment design.

Chapter 4 provides the design of the experiment and provides support to the decisions that lead to the implementation of this research.

Chapter 5 defines the analysis methods used to post-process the experimental data, as well as the analysis and insights gathered from the data.

Chapter 6 reflects on the results observed, their limitations and opportunities for future research.

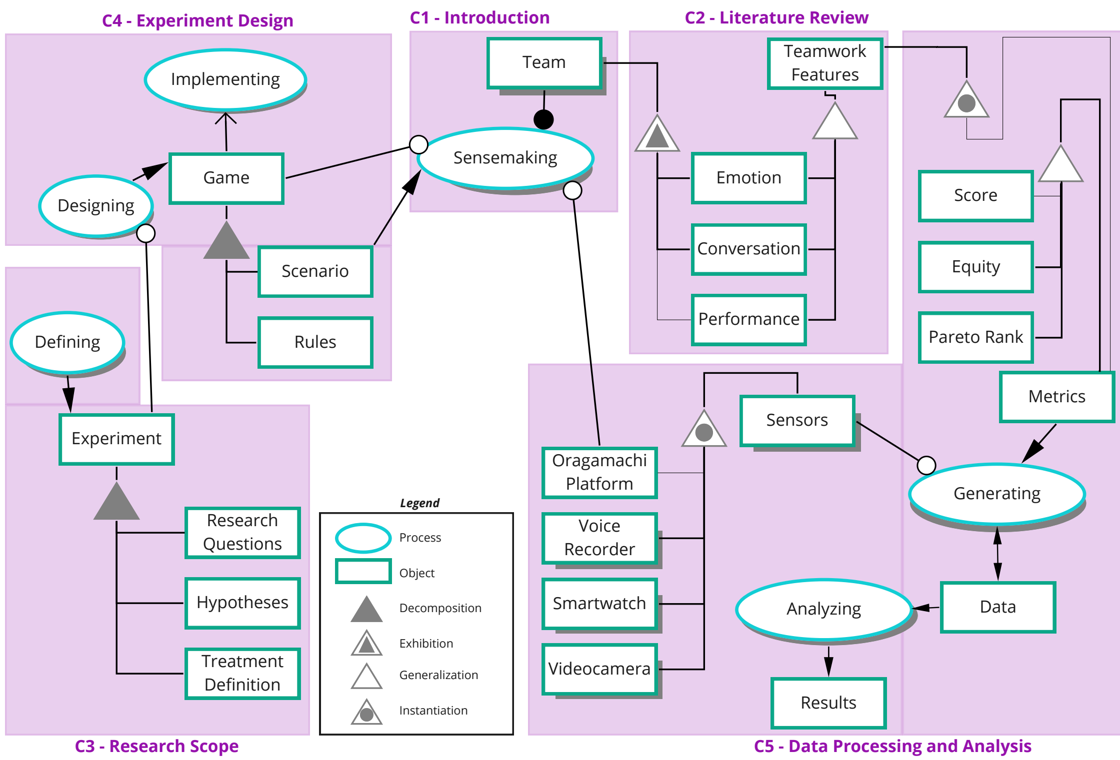


Figure 1-2: Thesis outline in OPD format.

Chapter 2

Literature Review

The cold never bothered me anyway.

Elsa, *Frozen* (2013).

Chapter 2 presents an overview of prior work in the fields of sensemaking (**Section 2.1**), social interaction measurements (**Section 2.2**) and teamwork (**Section 2.3**). It will also describe the context around the instrumentation of the sensemaking process and the position of this thesis with respect to it as well as additional Definitions.

A summary of the chapter is available in **Section 2.4** with key takeaways.

2.1 Sensemaking

How can people, teams, and organizations engage with complex and confusing situations? This is one of the questions at the heart of work in organizational psychology. The concept of sensemaking (see Definition 1) plays a prominent role by providing theories that aim to explain how individuals and groups attempt to interpret novel and ambiguous situations. There is a well-established body of literature discussing these theories and mechanisms [Russel et al., 1993], steps [Beal et al., 2005], properties [Weick, 1995]; all of which provides interesting frameworks to characterize sensemaking across scales and have captured cases in different contexts.

Since the research around sensemaking has been so diverse, researchers have also worked on finding underrepresented areas that are interesting to continue maturing the knowledge of this field [Maitlis, 2005]. Of specific interest are to explore new methods to capture the sensemaking process [Kozlowski & Chao, 2018], bridging the gap between individual- and group-level processes [Stigliani & Ravasi, 2012] and the role played by emotion [Maitlis, 2005].

To better understand the intersection of said areas of opportunities, it is necessary to look at a few cornerstone sensemaking theories. First, to be able to measure the sensemaking process, it is essential to know what it is that we are trying to capture. Weick's sensemaking properties [Weick et al., 2005] are a well-known framework to describe instances of sensemaking. The following list provides a summarized version:

1. **Sensemaking Organizes Flux**, where humans face chaotic situations in need for understanding.

2. **Sensemaking Starts with Noticing and Bracketing**, where perception, language and synthesis play a role in enacting a list of actionable options.

3. **Sensemaking is About Labeling**, where, in the face of novel challenges, humans name abstract groups of experiences to ease processing the information acquired by experience.

4. **Sensemaking is Retrospective**, were the relationship of disaggregate pieces of information is given meaning as new information and context are discovered.

5. **Sensemaking is About Presumption and Action**, where people create an expectation of future results aimed at taking action on the system.

6. **Sensemaking is Social and Systemic**, where people are not relying exclusively in their individual logic capabilities but in other people's expertise and interaction with the system.

7. **Sensemaking is About Organizing Through Communication**, were groups of individuals are needed to engage effectively with a system challenge.

These properties' abstract nature makes them a popular tool for evaluating processes at different scales and contexts. At the same time, it is what increases the difficulty of *measuring* the process. Recent exploratory work [Vazquez et al., 2022] proposes insights on bridging the gap between the abstract sensemaking properties and measurable events and team behaviors that can be used to characterize sensemaking across disciplines. One of the key observations is that properties related to (a) presumption and action, (b) the social and systemic aspects, and (c) organizing through communication have measurable manifestation through individual and team observable behaviors.

Since sensemaking is only meaningful in a complex environment, it is essential to distinguish between the actors involved and the system/problem with which they interact. Therefore, it is also necessary to understand how that interaction takes place. For starters, the system/problem ought to have *cues* accessible to the agents [Maitlis & Christianson, 2014], and they will usually be ambiguous with an uncertain relation to the outcome of the problem. At times, sensemaking will also be triggered by unexpected events [1], when a discrepancy is created between the agents *prior* expectations and the changes in the system/problem caused by the crisis.

For people to make sense of a situation, they need to be able to take action and pay attention to the cues generated by said action, especially since they are often thought to be plausible [Weick, 1988]. Action and cognition are thus recursively linked. Actions shape the environment for sensemaking.

Aside from the interactions between agents and systems, the level of aggregation (individual $\longrightarrow$ team $\longrightarrow$ organization) at which sensemaking takes place is also key ([Pirolli & Russell, 2011], [Batterman, 2011]). If sensemaking occurs in a person's head, then collective sensemaking in organizations becomes a process through which more influential individuals episodically persuade others to think as they do. The exercising of power has been attributed to either (a) personality traits of individuals [Cooren, 2004], as well as (b) expertise associated with the problem at

[1] A crisis might not always create the need for sensemaking. If the crisis causes a reaction that is too strong on humans, it can block cognitive processes at any level, from individual to organizational [Maitlis et al., 2013].

hand [Hardwig, 1985]. It is only reasonable to imagine the increasing relevance of expert teams' sensemaking processes in a world with complex problems working in an uncertain environment leading to the need to capture snapshots of agents as they engage with each other through time [Lynam & Fletcher, 2015].

2.1.1 The role of emotion

As has already been mentioned, negative and moderately intense felt emotions (see Definition 5) are most likely to signal the need for and provide the energy that fuels sensemaking in organizations [Maitlis et al., 2013]. Research has also identified additional factors, such as its role in detecting and attending to anomalies [Adler & Obstfeld, 2007] and influencing how critical cognitive and social processes are interpreted [Schwartz & Clore, 2007].

Definition 5 (Emotion)**.** A transient feeling state with an identified cause or target that can be expressed verbally or non-verbally [Grandey, 2008].

Especially relevant to the research work presented in this thesis is the reciprocal relationship between emotion and mental model evolution [Rafaeli et al., 2009] (see Definition 6). In this regard, the literature presents several perspectives on tension. First is the view that positive emotions are beneficial in creating and disseminating mental models within a group; where a team could achieve higher quality team models, these models would be shared to a greater extent among members [Rafaeli et al., 2009]. Second, the acknowledgment that emotionality will influence the mental models depending on the nature of the problem/system and the organizational context [Maitlis et al., 2013]. Third, the notion that mental models are not influenced as much by whether emotions are positive or negative but rather by how intensely [2] they are felt [Maitlis et al., 2013].

[2] At the same time, highly intense emotions (such as panic and rage) may forestall sensemaking processes because they tend to interrupt thought processes, consume capacity and redirect attention away from the triggering event to the emotion itself. ([Beal et al., 2005], [Loewenstein & Lerner, 2003], [Stein, 2004]).

Definition 6 (Mental Model)**.** Representation of some domain or situation that supports understanding, reasoning and prediction. They include our beliefs about the networks of causes and effects that describe how a system operates, along with the boundary of the model and the time horizon we consider relevant —our framing or articulation of a problem [Sternman, 2000].

Regarding which emotions are expected to be helpful, literature recognizes two relevant types of processes. First, *generative processes*, which at an individual level are similar to "resourceful" of "horizon-expanding" sensemaking processes in groups [Wright et al., 2000], such as brainstorming sessions. They are associated with positive emotions. Second, *integrative processes*, which leads to more precise constructions of a situation based on more critical analyses of new information [Schulman, 1993]. They are usually necessary when multiple stakeholders, with their corresponding differing mental models, ought to come together to solve a problem. They are associated with more stressful and negative emotions. Regardless of which type of sensemaking process might be present, there is reasonable agreement that sensemaking is an effortful, sometimes tricky, and potentially unpleasant process. So, individuals must be energized to engage in it ([Maitlis et al., 2013], [Russel et al., 1993]), leaving a door open to interpersonal dynamics.

The challenge with getting an improved picture of the role emotion plays in sensemaking is to measure it, especially when the level of aggregation is above one individual. At this point, it is essential to mention research work from the field of neurology. Even though arousal in the autonomic nervous system is not the same as the experience of emotion, research consistently links the two, with significant evidence that different patterns of autonomic arousal connect to different kinds of emotional experience [Levenson, 1992].

2.1.2 The role of collective work

Work from MIT's GTL has focused on how interactions across teams-of-teams can have a role in a team's efficacy in evaluating, proposing, and creating designs for complex systems ([Pelegrin et al., 2019], [Tan & Moser, 2019]). Although it has not been explicitly called as such in that body of research, the experiments are good examples of collective sensemaking (see Definition 7).

Definition 7 (Collective Sensemaking)**.** The process by which groups create shared understanding of a situation, and in particular, how individuals who hold different pieces of information are able to construct new meaning together [Weick et al., 2005].

This body of literature recognized that multiple materials and processes are involved in how groups and organizations create shared understanding. Examples are covered, ranging from software's User Interfaces [Russel et al., 1993], physical prototypes [Yang, 2005] and surgical instruments [Hindmarsh & Pilnick, 2007]. The effect of instruments on collective work becomes even more relevant once the *interaction* of material and conversational practices is studied [Stigliani & Ravasi, 2012]. A common point in the previous cases is the use of either open or semi-open problems/systems, which of course, allows for an understanding of sensemaking in a realistic environment but causes complications in isolating and controlling for external influences.

2.2 Measuring Social Interaction

Visualizing how a small group of people collaborate often relies on behavioral data such as speech data [Pentland, 2012], motion data [Sanchez et al., 2020], or psychological data [Gloor et al., 2021]. Because behavioral data can be measured unobtrusively in real-time, it is suitable for depicting the continuous change of teamwork as opposed to traditional survey-based methods for investigating teamwork [Kozlowski & Chao, 2018]. In either survey-based approach, the purpose is still to improve our understanding of the current team state and, ideally, indicate how it could be improved.

For example, the development of *Meeting Mediator* [Kim et al., 2008] can show the interactivity of teams and the participation of each member in real-time based on data from sociometric badges. At the same time, research has found that nonverbal data can indicate various constructs such as team roles [Vivian et al., 2015] and member influence [Basu et al., 2001].

Similar to the roles emotion (Subsection 2.1.1) and collaboration (Subsection 2.1.2) play in sensemaking, measuring social interaction is subject to the level of aggregation at which it is applied. At an organizational scale, instruments such as network visualizations built using email databases can show communication flows and patterns [Gloor et al., 2012]. A natural extension of said work is the comparison of the patterns across multiple contexts, such as different global communities. These long periods may be meaningful for large organizations, but their applicability is problem-dependent as with any other instrument. A team-level characterization of communication patterns might require different instruments ([Pentland, 2012], [Gaggiolo et al., 2020]).

An approach for studying fine-grained process data comes from recording sensemaking, which is accomplished in real-time. Several methods — including conversation analysis, discourse analysis, and especially micro-ethnography ([Liu & Maitlis, 2014], [Cooren, 2004]) — could be used to reveal how participants make sense during the strategizing process and strategizing process [Kaplan & Orlikowski], exploring both hidden qualities of the unfolding process and how it related to teamwork, coordination. ([Brown, 2000], [Maitlis & Christianson, 2014], [Cooren, 2004], [Hindmarsh & Pilnick, 2007]).

Recent work is starting to create methods such as multi-sensor measurements to capture the dynamics of individual movement and speech and social interaction at the team scale [Stefanini et al., 2020]. Similarly, there have been efforts to characterize team-level cognition of medical teams based on the change in real-time communication patterns [Gorman et al., 2020]. The research work presented in this thesis is, thus, adjacent to this field. Here, a concept is proposed that links the interaction between human-system interaction within a given context: a sensemaking pattern (see Definition 8).

Definition 8 (Sensemaking Pattern)**.** It is the overall behavior over time of the team socializing structure, the team sentiment and the evolution of system exploration/exploitation solution patterns.

2.3 Teamwork

Although small groups and teams have been studied with various research methods, the more commonly used methods — survey, experiment, and observation— are problematic for examining team process dynamics [Kozlowski & Chao, 2018].

The science of teamwork can go beyond self-reported data from surveys to access team processes and performance and to use more real-time methods, or "sensors," to measure the dynamic aspects of team processes and performance in a quantitative manner [Manandhar et al., 2020]. This area of research is expected to advance the science of teamwork to a new era beyond static models to a more dynamic view of teamwork [Mathieu et al., 2018].

As presented earlier in this chapter, much of the sensors research has been conducted using sociometric badges [Kim et al., 2012]. Data from these sociometric badges can be collected unobtrusively from many people over long periods at either low- or high-frequency rates. These data offer opportunities to research behaviors that are not readily available with traditional methods. However, the notion of transitioning to unobtrusive data capturing remains theoretical as the validation of the new instruments is only recently ramping up [Sanchez et al., 2020].

Leaving the social measurement aspect aside, work at MIT's GTL has focused not only on the interaction of multiple people working on models implemented in simulation software but also on examining the interactions among Social-, Problem- and Solution-Spaces [Fruehling & Moser, 2018]. Given the more controlled nature of workshop-based experiments, this approach could complement the need to validate specific team-level instruments. The rationale is that if the people are working on systems complex enough to require sensemaking and the researchers fully characterize the system, it can be possible to observe more controlled social effects.

This interpretation of teamwork relies on the assumption that most complex engineering problems are sociotechnical systems, as depicted in Figure 2-1.

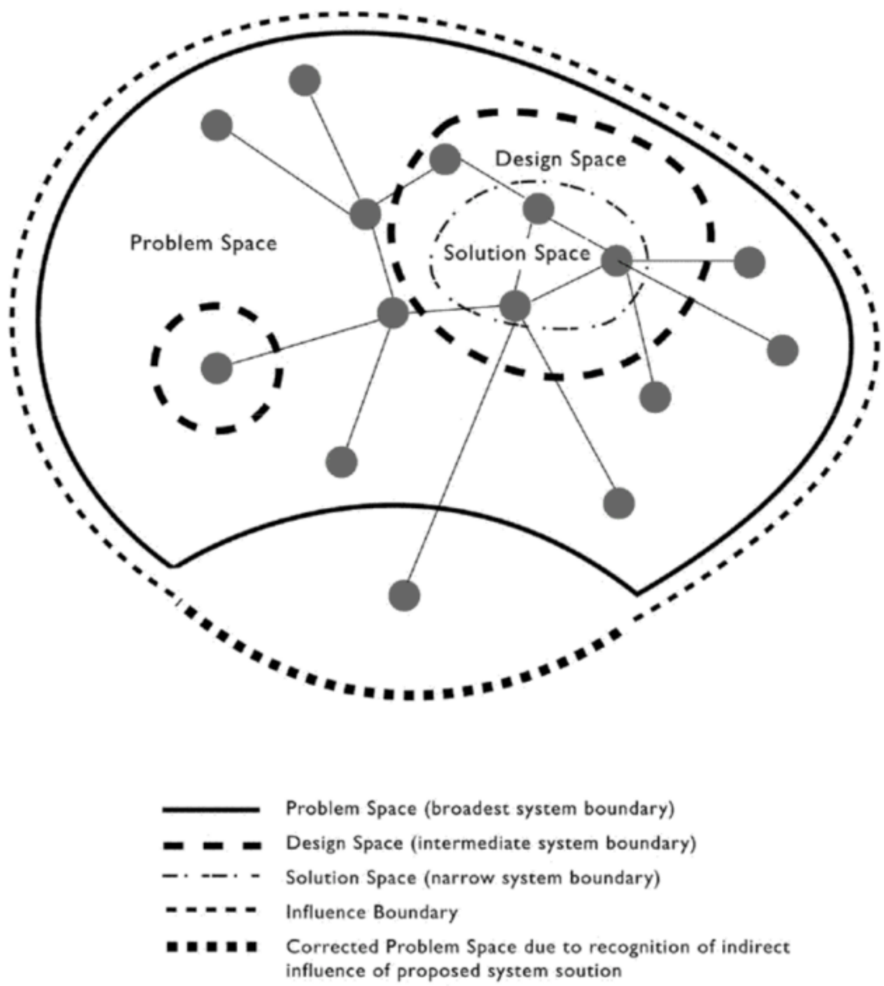


Figure 2-1: Design Spaces diagram. Representation of the complex relationships between the nature of problem (Definition 3), solution (Definition 4) and design (Definition 9) spaces due to sociotechnical interactions. Reproduced from [de Weck et al., 2011].

A fully characterized Design Space (see Definition 9) can then be visualized as a tradespace (see Definition 10) of Figures of Merit and the progression of a team plotted over it and evaluated as to have an objective measure of the team's performance over time. Thus, also enabling a gauge for the effectiveness of a team's learning journey [Pelegrin, 2018].

Definition 9 (Design Space)**.** Space of variables, from the system being acted upon, that are possible but might not be all feasible or desirable. It may include other solution elements that could have been part of the solution [de Weck et al., 2011].

Definition 10 (Tradespace)**.** It is the space spanned by the completely enumerated design variables, which means given a set of design variables, the tradespace is the space of possible options in the design space.

2.4 Summary

As was mentioned earlier in **Section 2.2**, most studies contributing to the understanding of teamwork and sensemaking are performed *in natura*. Additionally, the complex interactions happening between agents, system and context, make a comparison across contexts very difficult. This thesis aims to bring together aspects of sensemaking that have already been recognized by the research corpus as important yet difficult to integrate.

Chapter 3

Research Scope

We've been through worse.

Joy, *Inside Out* (2015).

This chapter serves a a linking ping between the research available in the field and the experimental design. Section 3.1 determines the overall objective of the research, its scope and most important constraint factors, the research questions addressed in this thesis and the resulting hypotheses.

3.1 Research Approach

The main opportunity area identified in **Chapter 2** is about gaining understanding on the process of sensemaking from an experimental standpoint. With the increasingly complex and ambiguous problems that dominate most engineering and scientific disciplines, the need for tools that support teams and organizations to design appropriate solutions will only become more pressing over time.

Although there are numerous instances of research characterizing sensemaking events in multiple organizations, there is still opportunity to create sensing instruments that are able to capture the team dynamics and their relationship to team performance. The added dimension of team-size and geographically-spread / asynchronous organizations, although important and interesting on its own, remains outside of the scope of this thesis.

3.1.1 Research Objective

The objective of this thesis is to provide an integrated sensemaking framework to characterize a team's (1) performance, (2) emotion journey, and (3)communication pattern. To achieve the objective, the thesis defines and applies a set of analytical instruments and a methods capable of capturing the sensemaking process for a simulated engineering design problem.

3.1.2 Research Scope

The experimental nature of this thesis requires a limited scope in terms of scale. A total of 39 participants joined a single two-hour stand-alone workshop. The teams worked independent of each other in a design for a Mars Colony using a virtual simulator. The teams were conformed by three participants and possessed no previous knowledge of the challenge.

The scope of the research was determined by numerous factors. Regarding the number of participants, the availability of sensors and space for the experiment were constraints. From the point of view of participants' availability, the workshops had to impose a minimum time requirement. Lastly, the complexity of the problem under analysis was determined by a trade-off between enough complexity so that a solution was not intuitive and low-enough that the system (Mars Colony) could still be fully understood and assessed by the researcher.

3.1.3 Research Questions

The questions formulated for this thesis start with defining assumptions found in the literature (see **Chapter 2**) and trying to test them out. Since the purpose of the thesis is on the instrumentation of the sensemaking process, it is reasonable to evaluate the methodology with situations that have wide support in the literature.

Chapter 3 provides an in-depth look at scientific background of the claims used in this thesis.

The first step is an evaluation of the role sentiment plays in a team's performance.

Assumption 1: The intensity of a team's sentiment influences its actions on the system being designed ([Rafaeli et al., 2009], [Maitlis et al., 2013]).

Then comes gaining understanding on the role of conversation patterns in a team and the overall performance [1].

Assumption 2: High performing teammates tend to have a balanced participation share among its members during conversations ([Kim et al., 2008], [Pentland, 2012]).

Correspondingly, the research questions (RQ) addressed in this thesis are:

RQ1. How does a team's emotional journey during teamwork influence the team's performance?

RQ2. How do conversation patterns influence team performance?

3.1.4 Hypothesis

From the literature review in **Chapter 2**, two main elements of team behavior, sentiment and communication patterns, were chosen to capture measurable changes in a team's sensemaking process. Given that there is evidence from neuropsychology that interaction between an individual's behavior and his or her cognition and emotion [Dolan, 2002] exist, by measuring both levels (1) individual and (2) team, a change in behavior can be used to represent cognition changes at team-level. A conceptual diagram representing this relationship is shown in Figure 3-1.

With this background, it is now possible to estate a hypothesis to evaluate Research Question RQ1:

[1] Measured by the progression of the system model metrics and how close they are to the targets provided to the team

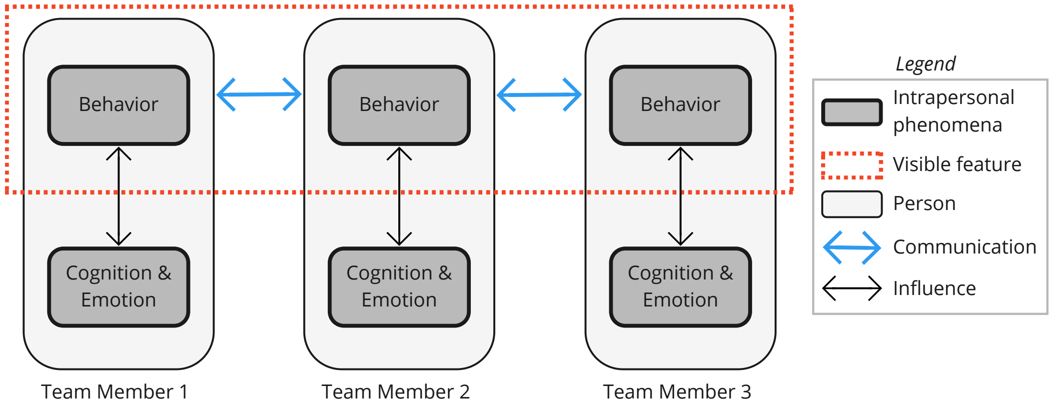


Figure 3-1: Individual behavior and individual states. Adapted from [Peng et al., 2022a]. It is worth pointing out in this diagram that the relationship between behavior, on a "surface" level and cognition/emotion is significantly more complex than a bidirectional line. The influence of the Affective-Behavioral-Cognitive model has also been explored for team dynamics [Ilgen et al., 2005]. For the purpose of this thesis, a presumption is made that the relationship of these two levels within an individual is strong enough that they can be used interchangeably.

Hypothesis 1 (H1)**.** Teams that experience a (1) more **intense** emotional journey and (2) **moderate frustration** during the process of sensemaking have a higher performance,

where **intensity** refers to magnitude of the sentiment expressed by the team members and **moderate frustration** refers to a heightened sensitivity of the team members to the (in-)consistency of new cues with respect to their emerging account [Maitlis et al., 2013] of the system model.

It is important to note that the relevance of frustration to sensemaking is caused by the following factors [Maitlis et al., 2013]:

1. Members of a group all poses different and conflicting information about their situation. Forcing the group to take an integrative approach to the creation of a single team mental model.

2. New information acquired by the group is not fully consistent over time —while still remaining plausible.

3. Human nature's proclivity to emphasize negative feelings and create meaning so they can regain control of its environment.

The other relevant aspect of sensemaking explored in this thesis is related to the role of verbal communication in the team. As has already been presented in **Chapter 2**, the relevance of communication in the sensemaking process has been explored extensively already. This thesis proposes a method to capture the communication pattern of a team and explore the way said pattern evolves as the team engages with a system model that is effectively a white-box.

The proposed hypothesis related to Research Question RQ2 is:

Hypothesis 2 (H2)**.** Teams that have an **equitable** and **stable** participation, among its members, during a work session have a higher performance,

where **equitable** refers to the time-share of each member's participation is roughly equal and **stable** means that it must not change as the work session progresses.

Even though this thesis does not dive into sensemaking patterns related to either (1) structural team differences or (2) system models that are significantly different, the experiment still captures data that be used in the future to characterize the participant population. More information is provided in **Section 4.3.1** on page 56.

Chapter 4

Experiment Design

If you are what you eat, then I only want to eat the good stuff.

Remy, *Ratatouille* (2007).

It is no coincidence that experimental teamwork research relies on close collaboration from multiple disciplines. **Chapter 4** takes a deep dive into the three main pillars needed to instrument the sensemaking process for a team. It will also continue providing necessary definitions to build a common language across the experiment.

First, **Section 4.1** presents a detailed description of the experiment design covering the theory behind the design decisions. It will also show the implications of the theory on the implementation of the experiment and how it was organized and executed. Secondly, **Section 4.2** since the experiment relies on a team playing a game, the design of it is also presented and discussed, together with an analysis of potential alternative game design options. This section will also provide insight into the implementation of the game. **Section 4.3** gives ample background regarding the instruments used to capture the sensemaking process on a team level.

A summary of the chapter is available in **Section 4.4** with key takeaways.

4.1 Experimental Design

The Global Teamwork Lab (GTL) is dedicated to developing methods and tools for unobtrusively instrumenting and analyzing teamwork in order to detect behaviors indicative of the phenomena that lead to high-performing teams ([Fruehling & Moser, 2018], [Manandhar, 2020], [Pelegrin et al., 2019], [Tan & Moser, 2019]). In order to achieve its mission, GTL needs to develop competencies that range from software development [Carroll, 2021] to the design of serious games [McDonough, 2021]. This characteristic makes every piece of work a collaborative effort in itself, usually brought to life by a small team with the indirect support of a small research community.

The basic structure of a *current* GTL experiment is as follows; (1) the people —a single team being the main unit of analysis and said team being conformed of three to five participants (see Definition 11), (2) the product [1]. — typically represented by a system model (see Definition 13) with which the team (see Definition 12) interacts. Previous system models studied by GTL have been: construction sites [Manandhar, 2020], V&V plans [Krehbiel, 2022] and. project management exercises [Fruehling & Moser, 2018]. Finally, (3) the process —a single self-contained workshop lasting between one and three hours, where each team works independently on a system.

Definition 11 (Participant)**.** Experimental subjects. People who join the experiment as subjects observed by the researcher a posteriori.

Definition 12 (Team)**.** A set of two or more individuals having specialized skills and diverse knowledge [2], that depends upon the actions and abilities of one another in order to achieve one or more shared goals.

Definition 13 (System Model)**.** Platform representing the dynamics of the product/technology on which the team is working.

---

[1] Specifically for this thesis, the system model is an Agent-Based-Model emulating the behavior of a population in a theoretical Mars Colony. More information about the model can be found in Subsection 4.3.3 on page 60.

[2] In GTL's context, the heterogeneity of the team is an important feature usually represented by a variety of roles and points-of-view.

The final ambition of GTL is to study more natural experiments where the systems-of-systems are worked on by teams-of-teams coming together. In most corporations, this would be 5 to 10 people representing different cross-functional teams coming together at critical junctures to make architectural decisions. However, the capacity of GTL up to this point in time has limited access to those natural conditions. The current experiment form —described in the previous paragraph— is still a work-in-progress that ought to mature to be able to address GTL's long-term mission.

4.1.1 Experiment Structure

The basic format of a GTL experiment has already been introduced at the beginning of this section (see page 40). The experimental design for this thesis mostly stays in the framework of the previous GTL work —which is from here on referred to as **Team Component**, and an additional **Individual Component**. All of the participants opt-in to join in the experiment.

The Individual Component is used to characterize the personality components of each participant. It relies on facial reaction, body movement, and physiological signals; all captured as the participant watches a recording with multiple video snippets expected to trigger specific emotions. Further information is presented in Subsection 4.3.1 on page 56.

The Team Component consists of a one-hour-long workshop where participants form randomly assigned teams of three people. Each team works independently of the other to interact with a system model. Before the experiment, the teams had no relevant information about the problem they would engage in. More details about the experiment sequence are provided in Subsection 4.1.3 on page 44.

The system model is provided in a dedicated workstation, and a single team member, selected by the team at the beginning of the session, is designated the *simulator* role. The role of *simulator* is important since it may have a disproportionate influence over the share of verbal communication said member has [3].

[3] As part of the experiment preparation, two dry-run sessions were executed. The purpose of these sessions is to test the readiness of the Research Team (see Definition 14), that all the instruments

Definition 14 (Research Team)**.** Students that actively contributed to the design of the experiment and instrumentation presented in this thesis. Qualified to make judgements about research methodology.

The definition of an experimental treatment that can influence teams' sensemaking patterns was particularly relevant to the Team Component of the workshop. Since Hypothesis 1 is related to the intensity of emotions, the expectation is that the treatment ought to create a different emotional pattern between groups. Simultaneously, the conversation pattern —related to Hypothesis 2— must be influenced by the treatment. The chosen mechanism to act as experiment treatment is to increase the stringency of the team's goal, making the required system performance much harder to achieve. The change in performance target is presented to half of the teams (Treatment Group), while the other half keeps the original targets. The news is delivered halfway through the session to familiarize the team with the simulation tool. In this way, two groups are established: **Control Group** and **Treatment Group**.

Chapter 2 presents multiple factors that, according to the reviewed literature, are likely to influence the way individuals and groups make sense of situations. Moreover, contingent on the number of participants for this experiment, it is essential to remove as many known factors that can alter sensemaking. Table 4.1 presents a list of the identified factors and the corresponding way it was dealt with for the Team Component of the experiment.

If the experiment treatment seems overly simplistic at this stage, rest assured that keeping things simple has been a lesson learned the hard way by GTL. The version of a teamwork experiment proposed in this thesis accounts for as many failure modes identified by previous GTL members. Instances such as (1) having too many treatments [Manandhar, 2020], too few participants [Krehbiel, 2022], not enough teamwork time [Pelegrin, 2018], too complex system model [Tan & Moser, 2019] or

work correctly in due time, and also serves as a sanity check for the experiment sequence. During the first dry-run on June 23rd, 2022, one of the participants provided feedback that he felt compelled to talk during the workshop because he was the one using the keyboard and mouse. This feedback was integrated into the experiment design by capturing in a document which team member is designated as *simulator*.

Factor	Description	Reference	How did was the experiment adapted?
Unexpected events	For an event to trigger event, the discrepancy between the team's expectations and what it experiences should be (1) great enough and (2) important enough. Too little or too much of either, or both, and sensemaking might not be necessary.	(Maitlis, et al., 2014)	The scenario is made self-contained. Teams are informed that there is no more information available and no other action outside of what the system model allows them to do.
Ability to act on the system	The team learns by receiving feedback on how their actions influence the system.	(Maitlis, et al., 2014)	The simulator provides numerical information a few seconds after the teams commits a potential system design.
Social Power Differences	Team members with a perception of each other's power that is too great can distort the overall behavior.	(Weick, et al., 2005)	Randomized team formation.

Table 4.1: Factors influencing sensemaking.

too open-ended team goals [Fruehling & Moser, 2018]. Many of these experiment pitfalls have been already collected and published [Pelegrin et al., 2018]

4.1.2 Team Formation and Mental Model Alignment

Experiment participants were assigned to randomly assigned teams to participate in the experiment. The combined time of the Individual and Team Components was two hours, of which 45 minutes correspond to the former and the rest of the time to the latter.

A team is conformed of three participants, and each one of them is assigned an individual role with specific information. The team as a whole is also provided with a team briefing. The samples of these materials can be found in Appendix A (Figure A-4).

The experiment conforms to the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). The corresponding guidelines were used to perform the experiment.

Since the number of teams for the experiment is not high, a single treatment that is presumed to affect the team's sensemaking approach is used. For this to be visible, it is critical that both groups (Control and Treatment) have the same starting point. At the beginning of the Team Component of the experiment, all the teams are given precisely the same information and have the same starting configuration in the system model. A detailed overview of the system model and its implementation as a User Interface is provided in Subsection 4.3.3.

Previous GTL research [McDonough, 2021] has already proposed a method for quantifying the alignment of mental models and awareness of systemic effect among a team of stakeholders exploring a systems model.

Given that the overall team objective is to meet a series of design constraints for the Mars site, they need to explore (see Definition 15) the tradespace of Star City using three design decisions; (1) Population Allocation, (2) Connection Mode Allocation and (3) Functional Space Activity Allocation.

Definition 15 (Exploration)**.** The act of a team using a computational model to experiment with attributes of that model in order to learn about the system the model represents.

Previous GTL experiments [McDonough, 2021] theorize that design teams can be expected to follow different patterns in the problem space (see Figure 4-1 and 4-2) depending on their ability to align their mental models and the system awareness [Manandhar et al., 2020]. Figures 4-1 and 4-2 provide an overview of the expected influence of a team's mental model alignment and awareness.

Each of these scenarios can effectively be a distinct sensemaking pattern, but instead of being only detected by looking at the tradespace walk (see Definition 16) but also by a team emotion and the communication patterns.

Definition 16 (Tradespace walk)**.** The sequence of system model configurations that describes the exploration pattern of a design team.

In contrast to previous experiments that gave different starting site designs to different teams or provided different information to the teams, in the beginning, this thesis provides a single starting design point to every team.

4.1.3 Implementation Aspects

This subsection describes the specific implementation of the workshops and elaborates on how and why certain decisions were made. It starts at a high level of aggregation:

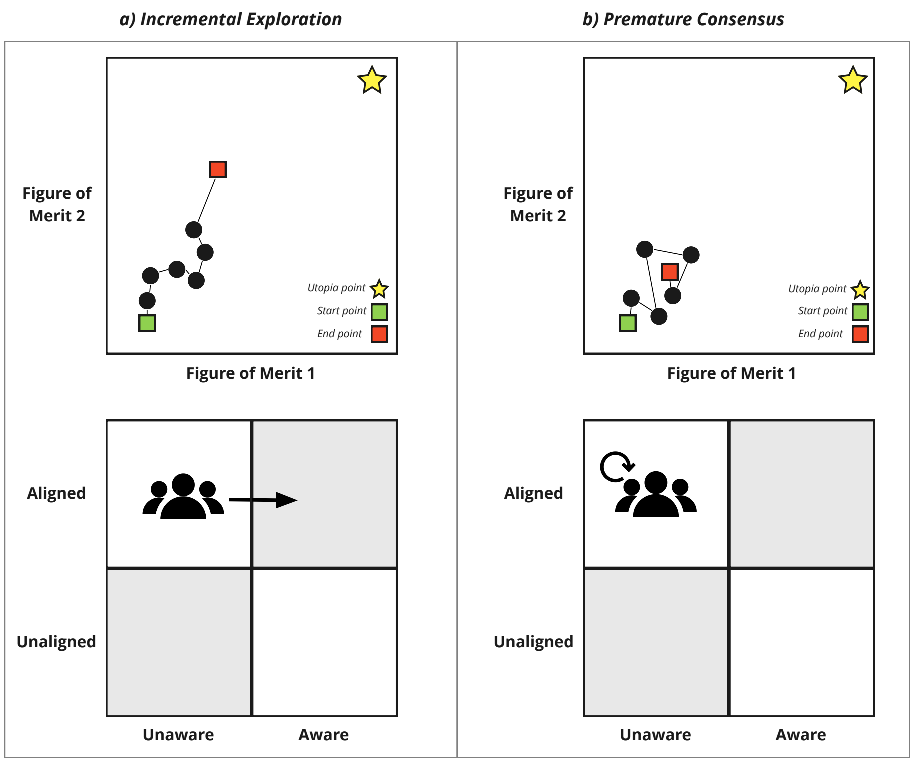


Figure 4-1: Notional effects of exploration diagrams 1. Adapted from [McDonough, 2021]. Scenario (a) **Incremental Exploration** would indicate a clear and progressive movement of the system model performance toward the Utopia point, driven by a team that, over time, aligns its members' mental models. Meanwhile, scenario (b) **Premature Consensus** shows a team whose members have a close mental model but do not become aware of it; in such a case, the team does not realize how to improve the system model performance structurally and gets stuck.

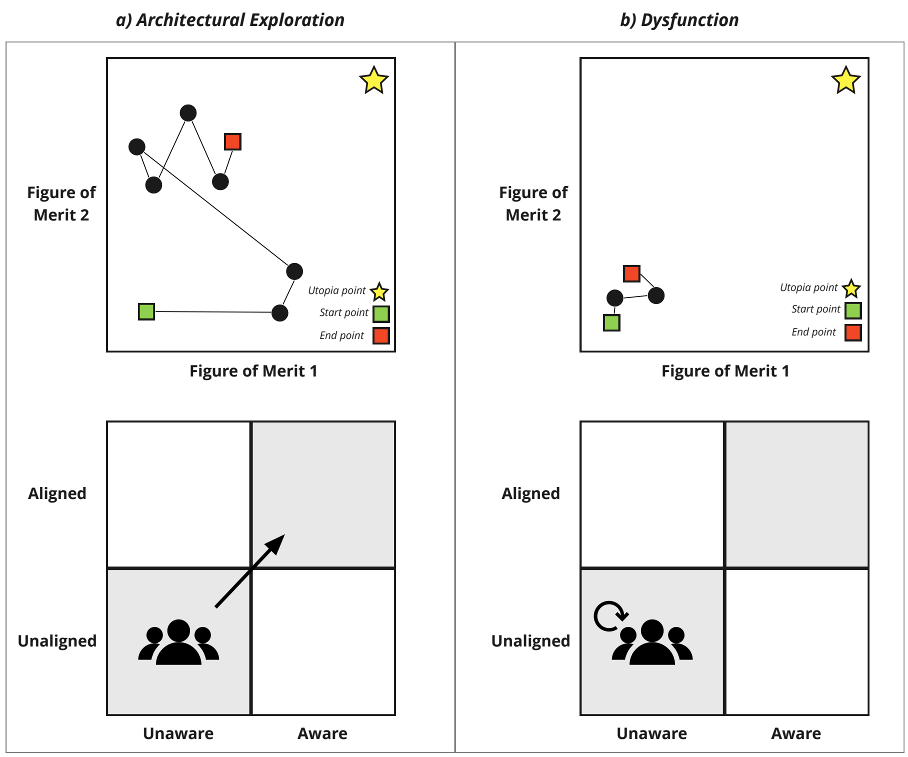


Figure 4-2: Notional effects of exploration diagrams 2. Adapted from [McDonough, 2021]. Scenario (a) **Architectural Exploration** effectively starts in a place of confusion for the team, and there are wild guesses as to how the system model reacts to the team's actions. A way to emulate the scenario can be to assign roles to the team members but provide no information about their value functions. Finally, scenario (b) **Dysfunction** represents a team that does not manage to understand the inner workings of the system and gives up.

Time / Day	Day 2	Day 3	Day 4
07h00 - 09h00	--	--	12 spaces
09h00 - 10h00	--	--	--
10h00 - 12h00	12 spaces	12 spaces	12 spaces
12h00 - 14h00	--	--	--
14h00 - 16h00	--	12 spaces	--

Table 4.2: Overview of the distribution of workshops during the week. *Note 1*: The numbering of the days in the table corresponds to the conference days. *Note 2:* The sessions were designed for 12 participants each, although only one of them turned out to be fully booked.

the week sequence and gradually refining the view, going to the look in a single day and finally, a single team.

The participants were a combination of attendees of a scientific conference during the summer of 2022. The experiment implementation was designed for 60 participants, but in the end, the actual participation was 39 people. The entire group was conformed of 25% female participants. The minimum and maximum ages were 17 and 60, respectively, with a group average of 35.4 years old. 65% of the participants did not have English as a native language.

Five two-hour-long independent sessions were planned over three days, and each session was fit for up to 12 individual participants [4]. See Table 4.2.

Since previous GTL research [Manandhar, 2020] indicates that team performance can be influenced by the time of the day at which the workshop is held, the day and time-block of each team are also recorded.

Control and Treatment groups were evenly split whenever possible for each one of the two-hour blocks. See Table 4.3. Whenever this rule could not be followed, the session Facilitators Team recorded the necessary changes (see Definition 17). In the same way, if due to either the lack or the excess of participants, changes had to be made to the number of people in a given team, these changes were recorded. All teams that participated in the experiment were complete.

Definition 17 (Facilitator Team)**.** Team in charge of preparing the instruments,

[4]The limit of participants is defined by two constraints; (1) the number of breakout rooms available for the three days of the experiment and (2) the amount of instrumentation equipment available.

Day	Time	Type	Team	Participants
Day 1	Morning	Control	1	3
Day 1	Morning	Treatment	2	3
Day 2	Morning	Control	3	3
Day 2	Morning	Treatment	4	3
Day 2	Afternoon	Control	5	3
Day 2	Afternoon	Treatment	6	3
Day 3	Morning	Control	7	3
Day 3	Morning	Treatment	8	3
Day 3	Morning	Control	9	3
Day 3	Morning	Treatment	10	3
Day 3	Afternoon	Control	11	3
Day 3	Afternoon	Treatment	12	3
Day 3	Afternoon	Control	13	3

Table 4.3: Control and Treatment distribution over the week. This table shows the *actual* number of participants and teams.

loading software, orienting participants and are in charge of running the experiment.

The reason to have each time block split between Control and Treatment is, as was mentioned already, the potential different behavior and performance between an *early morning team* and an *afternoon* team.

We are moving further into detail. The next implementation abstraction step is to look at the organization of a single time block, e.g., Tuesday morning. A two-hour time block is split into an Individual Section (45 minutes) and a Team Section (1 hour), with 15 minutes needed for task transitions and delays. See Table 4.4.

The point-of-view for a team is presented in Table 4.5.

In order to minimize the cross-team behavior, each team worked in a separate breakout room. Said room has a single computer and monitor, which the team will use to interact with the system model, and is equipped with a camera and microphone to record the session. The Facilitators Team helps the participants with sensors and timekeeping. Figure 4-3 shows a top-level view of a single breakout room.

<table>
<tr><th>Time [min]</th><th>Room 1</th><th>Room 2</th><th>Room 3</th><th>Room 4</th><th>Room 5</th></tr>
<tr><td>15</td><td>Part. 1</td><td>Part. 2</td><td>Part. 3</td><td>Part. 4</td><td></td></tr>
<tr><td>15</td><td>Part. 5</td><td>Part. 6</td><td>Part. 7</td><td>Part. 8</td><td>Part. 1-4</td></tr>
<tr><td>15</td><td>Part. 9</td><td>Part. 10</td><td>Part. 11</td><td>Part. 12</td><td>Part. 5-8</td></tr>
<tr><td>5</td><td colspan="4">Break</td><td></td></tr>
<tr><td>15</td><td rowspan="4">Part. 1-3</td><td rowspan="4">Part. 4-6</td><td></td><td></td><td>Part. 9-12</td></tr>
<tr><td>15</td><td rowspan="4">Part. 7-9</td><td rowspan="4">Part. 10-12</td><td></td></tr>
<tr><td>15</td><td></td></tr>
<tr><td>15</td><td></td></tr>
<tr><td>15</td><td></td><td></td><td></td></tr>
</table>

Activity Type
Empty
Individual
Intro+SW
Teamwork

Table 4.4: Breakdown of a single two-hour session. Note that Room 3 and Room 4 have a space right after the *Individual Component* of the experiment is over. This is the need to combine the individual participants into teams. Additionally, the instrumentation and software in the breakout rooms need to be updated from Individual to Team Component. Although it requires additional effort, the proposed sequence allows participants to show up late without disturbing the entire group.

<table>
<tr><th>Time [min]</th><th>Task</th><th>Measurement Instrument</th><th>Measured Signal</th></tr>
<tr><td>5</td><td>Instructions and Informed Consent</td><td>NA</td><td>NA</td></tr>
<tr><td rowspan="4">10</td><td rowspan="4">Watch movie snippets</td><td>Smartwatch</td><td>Physiological signals</td></tr>
<tr><td>Room video</td><td>Emotion detection</td></tr>
<tr><td>Voice recorder</td><td>Voice</td></tr>
<tr><td>Plant</td><td>Human Motion</td></tr>
</table>

<table>
<tr><th rowspan="2">Time [min]</th><th colspan="2">Task</th><th rowspan="2">Measurement Instrument</th><th rowspan="2">Measured Signal</th></tr>
<tr><th>Control</th><th>Treatment</th></tr>
<tr><td>15</td><td colspan="2">Instruction Video and Briefings</td><td>NA</td><td>NA</td></tr>
<tr><td>20</td><td colspan="2">Teamwork</td><td>1. Smartwatch
2. Room video
3. Voice Recorder
4. Plant</td><td>1. Physiological signals
2. Emotion detection (team average)
3. Conversation pattern
4. Human Motion</td></tr>
<tr><td>0</td><td>--</td><td>Treatment</td><td>NA</td><td>NA</td></tr>
<tr><td>20</td><td colspan="2">Teamwork</td><td>1. Smartwatch
2. Room video
3. Voice Recorder
4. Plant</td><td>1. Physiological signals
2. Emotion detection (team average)
3. Conversation pattern
4. Human Motion</td></tr>
<tr><td>5</td><td colspan="2">Exit Survey</td><td>Printed Survey</td><td>Mental Model update</td></tr>
</table>

Table 4.5: Team level experiment organization. The top table shows the *Individual Component* of the experiments, and the bottom table is the *Team Component*. Both levels number the measurement instruments used and the corresponding signals each instrument acquires. The *Team Component* also relies on colors to show whether a team has a treatment or not. Aside from the treatment, the rest of the sequence remains the same for both groups.

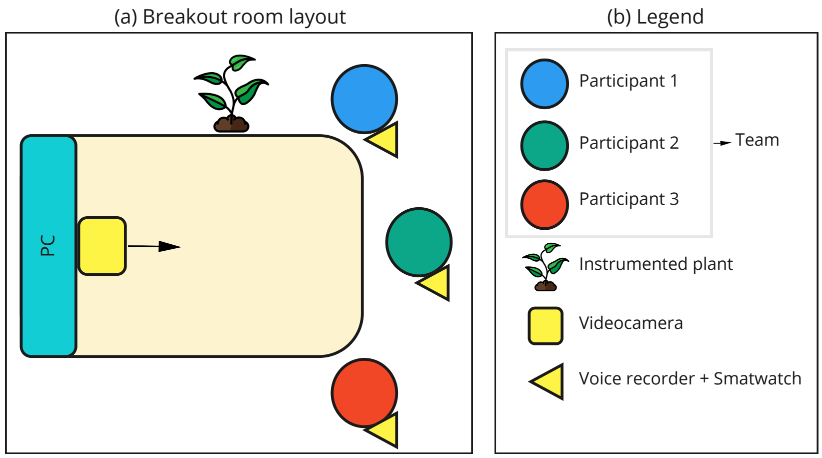


Figure 4-3: Layout for a single breakout room. There is a sample picture from one of the teams available in Appendix A-2.

4.2 Game Design

Designing a good game for an experimental setup is not a straightforward task. The presence and —to some extent— unpredictable participants' behavior require special care. Specific scientific fields, such as International Relations, have ample experience deploying mature games called wargames for social science research. Wargames are interactive events that display four characteristics; (1) human players, (2) immersed in scenarios, (3) bounded by rules, and (4) motivated by consequences [Lin-Greenberg et al., 2021]. These four factors then ought to be specified and described enough to facilitate possible future implementations of this experiment.

Although the design game used in this research is not intrinsically a wargame, the framework presented before is still relevant; given that it is also a type of serious game. Since it can generate confusion, the following definition will be used to refer to the game used in the research:

Definition 18 (Design Game)**.** An activity among two or more independent decision-makers seeking to achieve their objectives in some limiting context.[Abt, 1970]

Architectural Decision	Option 1	Option 2	Option 3
1. Game Type	*Experimental Game* *What is the effect of X on Y?*	*Observational Game* *How might decisionmakers behave in a specific type of event?*	
2. Participant Sample Type	*Experts* *RQ examines how a specific organization makes decisions. May have lower n of players*	*General Public* *Research Question examines general human decision-making behavior. Larger n of players*	
3. Game Play	*Realism* *Allow free-play.*	*Standardization* *Players select predetermined actions.*	
4. Moves	*One-off decisions* *Single move game.*	*Multiple decisions* *Multi-move game*	
5. Sides **How important is the other side's reaction to answering the research question?**	*Not important* *Single-sided game.*	*Important but seek to control outcomes of rounds* *1.5-sided game*	*Important but do not need to control outcomes of rounds* *2/multi-sided game*
6. Adjudication	*Low Control* *Use subject matter experts to adjudicate*	*Medium Control* *Use adjudication table*	*High Control* *Use fixed/predetermined outcomes*
7. Scenario Type	Highly Realistic Scenario E.g. real organization	*Abstract/general scenario* *E.g. fictional organization*	
8. Data Collection	*Motivation / Microfoundations* *Collect deliberative data*	*Game Outcomes*	*Both*
9. Unit of Analysis	*Individual*	*Team / Group*	*Game*

Table 4.6: Architectural decisions for game design. The selected options for the experimented presented in this research are indicated with a gray background. This morphological table is adapted from [Lin-Greenberg et al., 2021].

4.2.1 Design

Designing a game involves numerous decisions, many of which are relevant when the game is part of an experiment. To systematize the description of the game used in this thesis, a morphological matrix [Crawley et al., 2015] was created (see Table 4.6). It shows possible options for each relevant architectural decision (see Definition 19) in a Design Game. Several of the Architectural Decisions presented in the table have already been described earlier in this Section: (a) Game Type, (b) Participant Sample Type, and (c) Unit of Analysis. Except for Data Collection —which will be covered in **Section 4.3**, the rest will be explained in the following paragraphs.

Definition 19 (Architectural Decision)**.** The subset of design decisions that are most impactful in a system's design. They determine the performance envelope and encode the key trade-offs in a system [Crawley et al., 2015].

Game Play and Scenario Type

The experiment relies on a role-playing game where participants become stakeholders in a team aimed at finalizing the design of the first Mars Colony [5]. The motivation of having said scenario is to create; (1) a standard set of conditions for all the teams that suspend the reality outside of the game and hampers the influence of specific disciplinary experience and given individual participant may have, (2) provide the same starting point in terms of information and (3) create a basic expectation of how to use the system model given to the team (represented in the experiment by a simulation [Carroll, 2021]).

The Mars colony scenario sets up a shared goal within a fictional organization. The team's work is framed as being part of a company-wide mission run by the world's largest (fictional) aerospace company and setting their team goal to choose a site design for Star City's first settlement.

Primary Team Mission: Find a design that will lead to a successful Mars Colony.

Given that *mission needs* are implicitly shared through the goals and performance metrics of interest embedded in the individual stakeholder roles. Each role has an associated target (see Definition 20) Figures of Merit that link the problem space to their individual goals —provided in their briefing documents (see Appendix A). Suspension of disbelief [6] is achieved by establishing the imagined scenario 20 years in the future and clearly emphasizing that *no new technological development or validation is needed for mission success.* Meanwhile, specific knowledge related to the shared goal is provided in a Team Briefing document A-3.

Definition 20 (Target)**.** A specific aim that the stakeholder of a problem strives to achieve within a goal dimension. Targets are often expressed as "Key Performance

[5] The details of the colony are based on the work from MIT PhD students [Lordos & Lordos, 2019].

[6] The Oxford English Dictionary defines **(to) suspend disbelief** as *to refrain from being skeptical or from doubting the truth of something.* This concept is used frequently in serious gaming where it is needed for the participants to engage with a game as if it were real.

Indicators[7]" or as "Figures of Merit[8]".

From a practical standpoint, the team is asked to:

Team-level task: Use the provided system model to explore potential architectures and select your preferred design. Preference is decomposed into two evaluation aspects; (1) desirability (see Definition 21) and (2) feasibility (see Definition 22).

Definition 21 (Desirability (of a design))**.** The degree to which a design meets a stakeholder's individual and team goals.

Definition 22 (Feasibility (of a design))**.** Determination of a given system configuration being technically possible.

Moves

The game effectively allows a team to make one move by simulating their selected site design— at the time. The simulation tool computes the performance of the design and shows the results to the team in a tradespace, where the team can see how their choices compare to previous designs.

The game ends when time runs out.

Sides

The team is not playing against either the computer or other teams. It is not competitive in the traditional sense of *winning against an opponent*. This *single-sided* option is chosen because it allows for a cleaner comparison among Control and Treatment groups for the experiment.

However, previous research suggests that role-play has been an effective method for supporting learning by supplementing acquisition and use of knowledge through promoting affective engagement [Rooney et al., 2018]. In this workshop, role-play is used to establish context for the participants and provide them with sufficient motivation to realistically engage with the structures and conventions of a team and

[7] If related to project performance.

[8] If related to a Product's Technology performance.

		Stakeholder		
		Health Engineer	**Site Performance Engineer**	**Power Budget Engineer**
	Primary Goal	Ensure social engagement of sub-populations.	Ensure efficient use of the site.	Ensure effective use of power resources.
Value Function	**Ideally**	Diversity > 65 [%] Interaction > 3.5 [%]	Site Utilization > 32 [%]	Energy Use < 1,100 [kJ]
	Undesirable	Diversity < 55 [%] Interaction < 3 [%]	Site Utilization < 31 [%]	Energy Use > 1,500 [kJ]

Table 4.7: Breakdown of stakeholder goals and performance targets. Based on previous GTL work [McDonough, 2021]. *Note 1*: The indicated goal for Energy Use in this table corresponds to the Control Group. The Treatment Group's goal was 1.1 [MJ].

create an immersive social experience. Three relevant team roles were defined; (1) Health Engineer, (2) Power Budget Engineer and (3) Site Performance Engineer. An overview of their Goals, associated Value Functions, and Metrics are provided in Table 4.7.

Adjudication

Adjudication refers to the rules that dictate how human players interact with the scenario [Lin-Greenberg et al., 2021]. In this experiment, there is only one type of action the participants are allowed to perform to interact with the system model: simulate a design configuration.

4.2.2 Implementation Aspects

Game material with the individual and team briefings was prepared and printed by the Research Team before the experiments were run. The simulation software for the system model was also created and adapted before the experiments.

<table>
<tr><th>Sensemaking...</th><th>Signal</th><th>Instrument</th><th>Teamwork Phenomena</th></tr>
<tr><td>is about organizing through communication.</td><td rowspan="2">Conversation pattern</td><td rowspan="2">Voice Recorder</td><td rowspan="2">Team communication</td></tr>
<tr><td rowspan="4">is social and systemic.</td></tr>
<tr><td>Individual sentiment</td><td>Smartwatch</td><td rowspan="2">Sentiment</td></tr>
<tr><td>Team sentiment</td><td rowspan="2">Video Recorder</td></tr>
<tr><td>Personality Attributes</td><td>Individual Influence</td></tr>
<tr><td>is about action.</td><td>Simulation Runs</td><td>Oragamachi Platform</td><td>Systemic awareness</td></tr>
<tr><td>starts with noticing and bracketing.</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>is about labeling.</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>organizes flux.</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>is retrospective.</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>is about presumption.</td><td>NA</td><td>NA</td><td>NA</td></tr>
</table>

Table 4.8: Overview of instrumentation. The bottom cells indicated in red are not addressed in the experiment. Some insight on future research will be provided in **Chapter 6**. *Note 1*: This table still does not distinguish between individual and team-level signals. *Note 2*: The smartwatch instrument used to capture data during the experiment is shown in the table, but it was not used to perform analysis.

4.3 Instrument Design

The characterization of the process by which high-performing teams can achieve results has been an active research area, as has already been presented in **Chapter 2**. The importance of identifying teams and individuals that show a capacity to perform or shape people's behavior is highly valued by society at all organizational levels, e.g., companies and government. Nonetheless, the process of actually identifying said teams is complex. Previous research has tried to assess an individual's creativity with structural properties like the Torrance Tests of Creative Thinking [Sun et al., 2020]. This type of *static approach* to identifying high-performing teams can be extended by looking at *ongoing* activities involving problem-solving. The question then is, what phenomena should be measured and what are (un-)desirable patterns of behavior?

This Section elaborates on the tools and methods used to capture different aspects of teams' sensemaking process. Table 4.8 shows an overview of Weick's sensemaking properties studies in this thesis, the associated signals, and the instruments used to detect them.

4.3.1 Detecting Emotion

As was briefly mentioned in Subsection 3.1.4, the chosen approach to detect individual emotion is through behavior. As a continuation of previous efforts to capture the transitive properties of teamwork using an intelligent system to measure the workload and surgical performance of minimally invasive surgeons. [Sanchez et al., 2020], this thesis proposes an approach to measure the way used by teams to communicate and collaborate [Hindmarsh & Pilnick, 2007].

At the individual level, two instruments are used; (a) smartwatches and (b) facial recognition —only during the Individual Component of the experiment.

Smartwatches

The smartwatch [9] collects body movements through (1) accelerometer sensors, (2) heartbeat, (3) speech parameters [10] location changes through (4) GPS [11]. The data is stored locally in the watches during the workshop and must be exported at the end [12] [Sanchez et al., 2020].

The main measurements used from the smartwatches are an accelerometer sensor which measures the acceleration forces in m/s^2 on three physical axes, namely: x, y, and z. Thus, when the user moved the watch, e.g., shaking or tilting it, an acceleration to one of the given axes was caused. Second, a heart rate sensor collected the heart signals in beats per minute. The Happimeter App is used in *workshop mode*, which captures measurement constantly for the duration of the workshop [Roessler & Gloor, 2020].

[9] The watch is a **Ticwatch E3** smartwatch with Google's WearOS software. In order to capture the desired data, the smartwatches use an app developed by MIT's Center for Collective Intelligence: Happimeter `https://www.happimeter.org/`.

[10] Not used in this thesis. Instead, the experiment uses different voice recorders.

[11] Not used in this thesis since the teams do not change locations.

[12] This *temporal storage* feature was a modification done to the smartwatch software after the first experiment dry-run. The original configuration of the watches constantly transmits the data to the server. However, as is the case in this experiment, when many watches simultaneously send data, the server suffered transmission bandwidth errors and stopped recording the data, which led to losses that would impede analysis.

Facial Recognition

The relationship between facial emotion recognition and four well-established frameworks assessing different facets of personality; (1) Neo-FFI [Costa & McCrae, 2008], (2) Schwartz Moral Values [Schwartz & Bilsky, 1987], (3) Attitudes towards taking risk [Blais & Weber, 2006], (4) Moral Foundations [Graham et al., 2013]. Prior work has been done to characterize these frameworks by having an individual participant watch a recording design to create an emotional reaction while an algorithm evaluates the facial expression of the participant [Gloor et al., 2021].

A brief description of all the personality characteristics represented by the four frameworks is shown in Table 4.9. Since the teams in this thesis are assigned randomly, the motivation to capture individual personality traits is to have a snapshot of the participants. This snapshot can be used when the experiment is reproduced with a different group of people, e.g., in a different country.

At the individual level and specifically for this thesis, some factors may influence the conversation patterns. Therefore, during the Individual Component of the experiment, the *emotion reaction* video is used to characterize the individual personality traits of the participants. Table 4.10 shows a list of the snippets included in the video.

The same principle that is used at the individual level is used at the team level; the tool used to capture the team sentiment [13] has two main differences; (1) there is no emotional reference for the team as a unit —no predefined set of videos—, it is instead evaluated only by their reaction to the system model, and (2) the team is evaluated as a unit, so the sentiment level is an **average** of that showed by the team members.

Figure 4-4 and Figure 4-5 show a sample of the signals captured by Moody. It is worth mentioning that for the experiment in this thesis, there is a single Moody session that runs for a team, and it relies upon a single camera available in the breakout room.

[13] The tool is called *Moody*, `https://www.moody.digital/`. A student group from MIT's Collaborative Innovation Networks seminar —`https://sites.google.com/view/coinseminar22/home` developed the tool. Moody looks closely at the relationship between emotion through facial expressions and voice in presentations. Measuring and storing both facial and vocal emotions in real-time during live presentations with the help of Convolutional Neural Networks and trained emotion recognition models. [Page et al., 2021]

Framework	Factor	Description
Neo FFI Costa	**Neuroticism**	Disposition to experience negative affects, including anger, anxiety, self-consciousness, irritability, and depression.
	Extraversion	Tendency to focus on gratification obtained from outside the self.
	Openness	Receptivity to new ideas and experiences.
	Agreeableness	Ability to put others needs before one's own.
	Conscientiousness	Tendency to be responsible, organized, hard-working, goal-directed, and to adhere to norms and rules.
Moral Values Schwartz	**Power**	Social status and prestige, control or dominance over people and resources.
	Achievement	Personal success through demonstrating competence according to social standards.
	Hedonism	Pleasure or sensuous gratification for oneself.
	Stimulation	Excitement, novelty, and challenge in life.
	Self-direction	Independent thought and action--choosing, creating, exploring.
	Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature.
	Benevolence	Preserving and enhancing the welfare of those with whom one is in frequent personal contact.
	Tradition	Respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides.
	Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.
	Security	Safety, harmony, and stability of society, of relationships, and of self.
Risk Profile Blais	**Ethical**	Assess both conventional risk attitudes (defined as te reported level of risk taking) and perceived-risk attitudes (defined as the willingness to engage in a risky activity as a function of its perceived riskiness).
	Financial	
	Health / Safety	
	Recreational	
	Social	
Moral Foundation Graham	**Care / Harm**	Connecting perceptions of suffering with motivations to care, nurture, and protect.
	Fairness / Cheating	Sensitivity to evidence of cheating and cooperation, and reacting to emotions that compel one to play fair.
	Loyalty / Betrayal	Sensitivity to the need of forming cohesive coalitions.
	Authority / Subversion	Experience to navigate hierarchies effectively and forge beneficial relationships upwards and downwards.
	Sanctity / Degradation	Experience to develop a more effective "behavioral immune system".

Table 4.9: Personality traits overview. Note: The Risk Profile factors have the added nuance of capturing; (1) the perceived risk and (2) the taken risks for any given individual. This information is not presented here to simplify the table. Additional information can be found in [Blais & Weber, 2006].

Video Number	Short Description
1	Puppies -- Cute puppies running.
2	Avocado -- A toddler holding an avocado.
3	Condom ad -- Child throwing a tantrum in a supermarket.
4	Runner -- Competitive runners supporting a girl from another team over the finish line.
5	Maggot -- A guy eating a maggot.
6	Soldier -- Soldiers at battle.
7	Trump -- Donald Trump talking about Mexican mass migration.
8	Mountain Bike -- Mountain biker on daring ride down a rock bridge.
9	Roof Bike -- Guy biking on top of a skyscraper.
10	Roof Run -- Guy balancing and almost falling on top of skyscraper.
11	Racoon -- Man beating a racoon to death.
12	Abandoned -- Social worker feeding a starved abandoned black toddler.
13	Waste -- Residents collecting electronic waste in the slums of Accra.
14	Dog -- Sad dog on the gravestone of his master, missing him.
15	Monster -- Man discovering an invisible monster through the picture on his instant camera.

Table 4.10: List of video snippets used for personality assessment [Gloor et al., 2021].

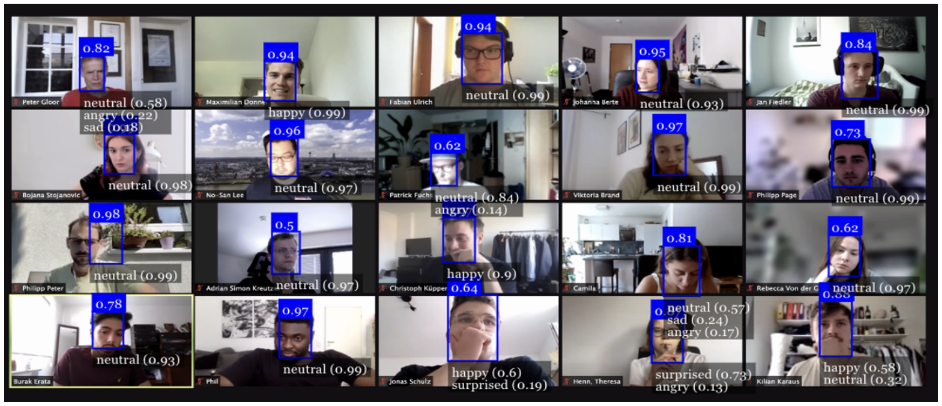


Figure 4-4: Presenter's view of the detected faces in an online meeting. Taken from [Page et al., 2021].

The data stored in the Moody server can also store the period-to-period *Emotion Radar* composition, which can give an overview of the dominant emotions over time. The Emotion Radar recognizes neutral, surprised, happy, fearful, disgusted, angry, and sad emotions based on the audience's facial expressions. However, even though these states have been recognized as descriptive of static captures of people's emotions —akin to categorical emotions— [Keltner et al., 1997], Moody uses a Computational Neural Network (CNN) implementation to recognize emotional states. The training database [14] for Moody does not recognize every possible emotion or combination thereof; thus, a *neutral* emotion was added to the model.

It has already been recognized in CNN-based recognition of emotions via facial expressions that the training databases are purposefully clear-cut and use professional actors. An environment such as a work meeting is prone to being formal and informative, which can lead to a neutral category being recognized by automated systems [Keltner et al., 1997].

4.3.2 Detecting Communication Patterns

The primary assumption for team communication in this thesis is that the change of communication patterns over time [Peng et al., 2022a] is relevant to capturing how teams make sense of a complex problem when the requirements change and become

[14]RAVDESS, TESS, JL-Corpus, and EMO-DB.

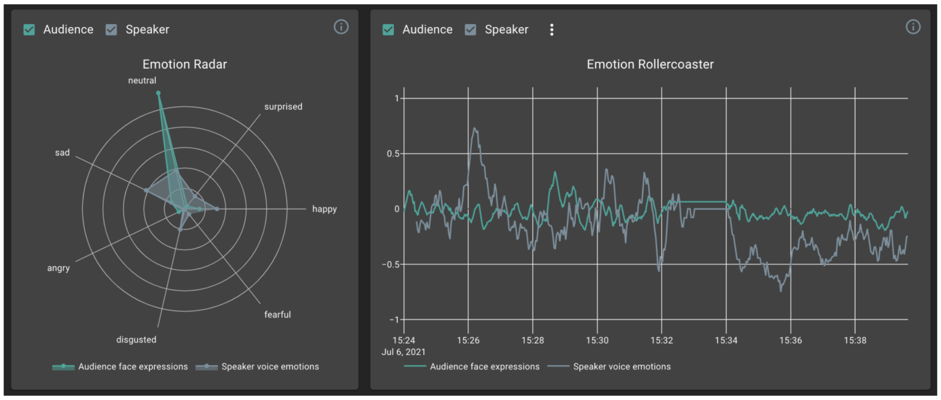


Figure 4-5: Sample of Moody data. Recorded statistics of the status during an online presentation. Taken from [Page et al., 2021].

more stringent. A similar association has already been studied, linking communication patterns to team learning [Peng et al., 2022b]. In their case, the learning gains were measured by the rate of students who created ideas with structural similarity to analogy cases. However, creating appropriate ideas might not necessarily reflect the understanding of analogical thinking.

Even though the application of physiological synchrony is a promising indicator for representing the real-time state of a team, its application in organizational science is minimal [Peng et al., 2022c]. Moreover, it is likely insufficient to capture nuances between individual and team-level reactions. Therefore, in this thesis, both instruments are used; physiology data from the smartwatches and communication patterns via the voice recorders.

4.3.3 Detecting Team Performance

One of the purposes of using a simulation model is to eliminate the subjectivity of team performance. It is not that being subjective is wrong, but the goal here is to reduce the noise so that we can isolate the contribution of team communication, emotion, and their influence over the system performance. This research consciously takes a different approach to studies that rely on judges [Sosa & Marle, 2013].

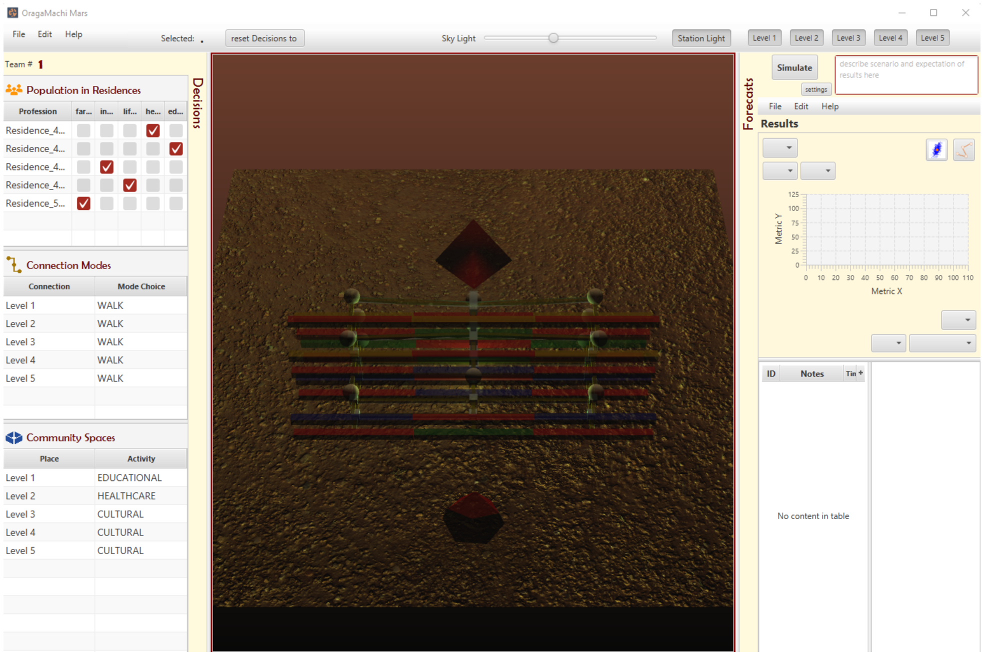


Figure 4-6: Mars Simulator user interface.

System Model - The Oragamachi Platform

In this experiment, a system model of the Mars Colony was used. The system/problem is an adaptation of the GTL's Oragamachi platform ([Carroll, 2021], [Moser et al., 2019]). The system model is a simplified version of that used by [McDonough, 2021]. The problem was designed so that it can be easily understood by general non-engineering participants in a short time —about five minutes. The problem description was provided in a single-pager and a facilitator briefing. The instructions to use the system model were provided as a video recording.

Teams were instructed to discuss and find a solution that considers the needs of all internal stakeholders. A section of the system model Graphical User Interface (GUI) (see Figure 4-6) functions as a scratch-pad where teams provide notes and a rationale that can be analyzed during the experiment's post-processing.

The starting design with which all teams start is presented in Table 4.11. All these are shown in the User Interface as options. This specific design has a good performance for a couple of metrics but has enough space for improvement for the team to —realistically— explore within one hour.

Decision	Factor	Choice
Population in Residences	Farmer	Level 5 Center Middle
	Industry	Level 4 Top Left
	LifeSupport	Level 4 Top Right
	Health	Level 4 Bottom Left
	Educator	Level 4 Bottom Right
Connection Modes	Level 1	Walk
	Level 2	Walk
	Level 3	Walk
	Level 4	Walk
	Level 5	Walk
Community Spaces	Level 1	Educational
	Level 2	Healthcare
	Level 3	Cultural
	Level 4	Cultural
	Level 5	Cultural

Table 4.11: Initial Mars Colony Configuration.

Site Evaluation Metric	Unit	Definition
Site Utilization	[%]	The percentage of spaces within the site with people actively engaging in a functional space activity at a given time.
Energy Use	[kJ]	The amount of energy being consumed by people moving through the site and engaging in activities.
Average Interaction Probability	[%]	The measure of the likelihood two people within a space will serendipitously "interact" due to their proximity to each other.
Average Diversity Score	[%]	The relative measure of population diversity within spaces on the site.

Table 4.12: Mars Site Evaluation Metrics.

The simulation output includes all the resulting metrics the team must evaluate, see Table 4.12. As mentioned, the participants will have been assigned specific metrics according to their roles and targets.

In this experiment, only three design decisions are available for the participants. Each design decision is limited to three options and can be taken for every level of the Mars site. In brief, they are:

- **Population Allocation Decision**: Which population groups reside on what level(s)?
- **Connection Transit Mode Selection**: Which modes of transportation should be supported in the long connections attached to each elevator lobby?

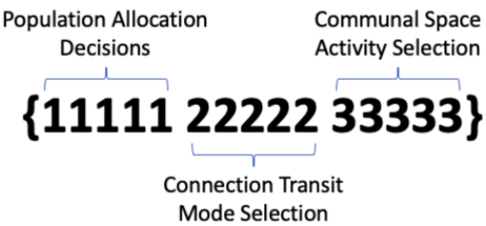


Figure 4-7: Star City design vector breakdown.

- **Communal Space Activity Selection**: Where should educational, healthcare, and cultural spaces be located on the site?

Each decision has five levels, and each of those has three options. So, an entire configuration can be defined in a set of fifteen digits, as shown in Figure 4-7. In this vector decomposition, the first five digits are related to Population Allocation, the second set to Connection Transit Mode, and the last set to Communal Space Activity. Each digit can be in the $1, 2, 3$ set.

The size of the full factorial (N_{FF}), considering a number of levels $l = 3$ and a number of factors $n = 15$, is computed using Equation 4.1 as follows,

$$\begin{aligned} N_{FF} &= l^n \\ &= 3^{15} = 14348907 \approx 14e6. \end{aligned} \tag{4.1}$$

Since it is not advisable to compute every possible configuration to generate a basis tradespace, a random sample of 1000 configurations was created and processed using the Oragamachi platform. Table 4.13 describes the input parameters used and their relationship to the design factors as seen by the participants.

The metrics' results can then be used to construct tradespaces in-lieu of a full-factorial solution space. Figure 4-8 shows the tradespaces for possible metric combinations and having *Energy Use* in common. Even though there are other possible combinations of metrics, such as *Interaction Probability* vs. *Diversity Score*, simple correlation analysis shows that combinations of metrics not including *Energy Use* do not have a strong trade-off. A graphic representation is presented in Appendix B-2.

Decision Set	ID	Description
Population Allocation	1	Residence on Level 4
	2	Residence on Level 5
	3	Residence split evenly 50/50 on Levels 4 and 5
Connection Transit Mode	1	Walk
	2	Electric Scooter
	3	Moving Walkway
Communal Space Activity	1	Educational
	2	Healthcare
	3	Cultural

Population Allocation Decisions					Connection Transit Mode Selection					Communal Space Activity Selection				
Farming	*Industry*	*Life Support*	*Medical*	*Education*	*Level 1*	*Level 2*	*Level 3*	*Level 4*	*Level 5*	*Level 1*	*Level 2*	*Level 3*	*Level 4*	*Level 5*
1	1	1	3	2	3	1	1	1	1	3	2	1	3	1
2	1	2	2	3	3	2	2	2	2	3	3	1	2	2
3	3	1	1	1	1	1	2	2	3	1	1	2	2	2
1	1	1	1	2	1	2	2	3	2	3	2	2	1	2
2	3	3	3	3	2	3	2	2	2	2	3	1	2	1
2	2	1	3	3	3	2	2	2	2	2	3	1	1	1
1	1	1	3	3	1	3	2	2	2	1	1	3	3	2
1	3	1	1	3	3	1	1	3	1	2	2	1	1	2
1	1	1	3	3	2	3	1	3	3	2	3	3	1	2
1	1	2	2	3	3	2	3	2	3	3	3	3	1	2

Table 4.13: Sample of Star City configurations. *Top table*: It shows the options for each of the three design decisions. *Bottom table*: It shows a sample of 10 random configurations. This format is also used in the **.csv* file to simulate within the Oragamachi platform. The top two lines in the table must not be used in the **.csv* file. Based on [Carroll, 2021].

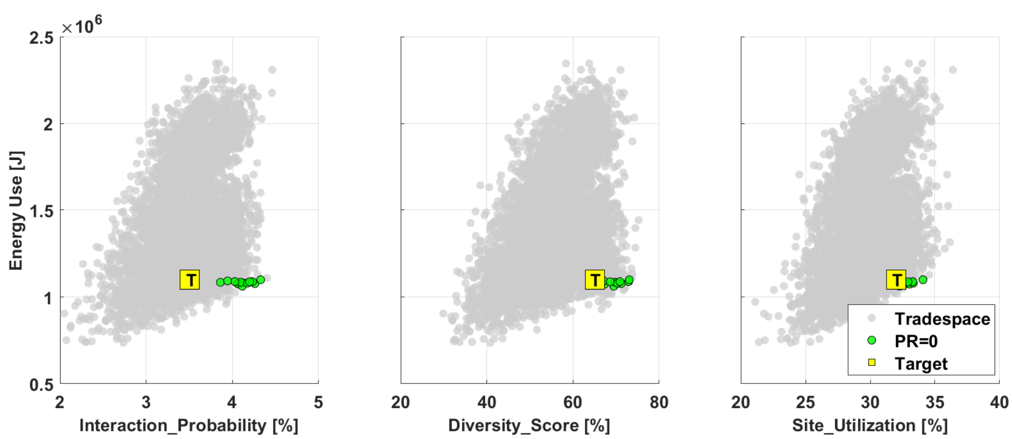


Figure 4-8: Star City tradespace sample overview. The x-axis in all three plots is expressed as a percentage. The y-axis is *Energy Use [J]* for all three plots. Each one of the gray dots represents a randomly selected unique configuration. *Note 1*: The "Target" yellow square represents the point in each tradespace with the location of requirement limits. *Note 2*: The green dots are locations of configurations that meet the requirements of the experiment. More information about the requirements and desirability of configurations is provided in Subsection 4.2.1.

4.4 Summary

This chapter presented the general approach followed by this thesis to integrate measurable aspects of sensemaking at the team level. As has been shown, experiments in the context of teamwork are not only to be created with the science in mind but also as a sort of *game*. The participants need to find an intrinsic motivation to participate and, for a moment forget that they are in the middle of an experiment. Implementing such an approach is a critical step, and the chapter presented guidelines and tools used during the workshops to capture data. Finally, an overview of the instruments to measure social behavior was introduced, together with the software used as the central game platform.

Chapter 5

Data Processing and Analysis

This isn't flying. This is falling, with style.

Buzz, *Toy Story* (1995).

Chapter 5 breaks down the process of evaluating the data from the experiment and assessing the results. In order to provide a meaningful evaluation of the results, the internal organization of the chapter is as follows: **Section 5.1** takes the data from a single team as a working example and shows how the team's (1) performance, (2) emotion, and (3) conversation pattern are evaluated. Afterwards, **Section 5.2** provides the linking between the team metrics and research questions RQ1 and RQ2, as well as the evaluation of both hypotheses 1 and 2. Finally, **Section 5.3** closes the chapter with a few key takeaways.

5.1 Analysis Approach

This section will take a single team's data source as a reference to lay out the evaluation process. Each of the Subsections will deal with a different team aspect: (1) performance, which is assessed based on the simulation results generated during the Team Component section of the experiment, (2) emotion, estimated using Moody and (3) communication pattern, captured using individual voice recorders. Methodologically, each of the three components was evaluated independently of each other

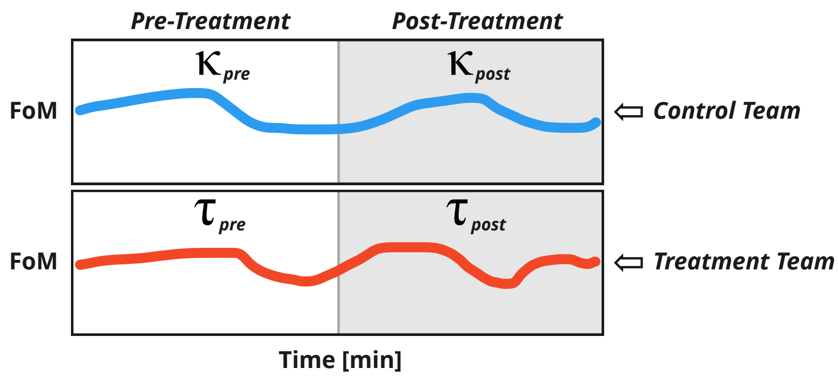


Figure 5-1: Breakdown of data sections for analysis. The choice of Greek letters to denote the different sections was based only on brevity, and it is easy to remember to which team type they correspond.

by individual researchers [1]

Before the start of the result evaluation, it is worth clarifying some nomenclature that will be used ahead. The experiment methodology detailed in **Chapter 4** estates two types of groups: Control and Treatment. It also explains the rationale behind the choice of a single-treatment experiment halfway through the team workshop. The decomposition of relevant sections to this research is represented conceptually in Figure 5-1. The plot shows the progression over time of a single metric for two separate teams and a gray line splitting pre- and post-treatment regions. Even though Control Teams were not subject to any treatment, the nomenclature proposed in the figure to distinguish the different regions will be used for consistency.

For some of the following (sub-)sections, it is important to be aware of whether the analysis is taking an isolated team (e.g. κ) longitudinal comparison (e.g. $\kappa_{pre} \longrightarrow \kappa_{post}$), cross-sectional comparison (e.g. $\kappa_{pre} \longrightarrow \tau_{pre}$) or cross-team comparison (e.g. $\kappa \longrightarrow \tau$).

5.1.1 Estimating Team Performance

The system model described earlier in Subsection 4.3.3 on page 61 mentions that any evaluation of the system involves multiple metrics and their corresponding ob-

[1]Collaboration with MIT's Center for Collective Intelligence – Jakob Krusse, and University of Tokyo – Sixiong Peng.

jectives. These problems are known as multi-objective optimizations, where any proposed "good" solution has to perform well in all dimensions. Multiple objectives create trade-offs between them where it is possible to improve upon one objective by choosing a less favorable position in at least one other objective.

Pareto optimality and ranks

This section's point of view is:

$\mathcal{K}$ - Single Team

A multi-objective optimization problem can be represented mathematically as a maximization or minimization problem that maps the input vector space [2] (x, p) to the output vector space [3] J subject to the constraints g and h:

$$\begin{aligned} \max \quad & \mathbf{J}(\mathbf{x},\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{g}(\mathbf{x},\mathbf{p}) \leq 0 \\ & \mathbf{h}(\mathbf{x},\mathbf{p}) = 0 \end{aligned} \tag{5.1}$$

Considering two feasible objective vectors $\mathbf{J}^1, \mathbf{J}^2$. $\mathbf{J}^1$ weakly dominates $\mathbf{J}^2$ iff:

$$J_i^1 \geq J_i^2 \;\forall\; i, \tag{5.2}$$

and

$$J_i^1 > J_i^2 \text{ for at least one } i. \tag{5.3}$$

The resulting Pareto front consists of a set of non-dominated solutions. For example, two output Figures of Merit were chosen with the Star City site, and their Pareto front was computed using the previous method. The result is presented in the left-side plot of Figure 5-2. The Pareto Front shown in the plot is the result of minimizing *Energy Use* while maximizing *Interaction Probability*.

The next step in the challenge to assess the team's performance is to evaluate

[2] Solution space.
[3] Problem Space

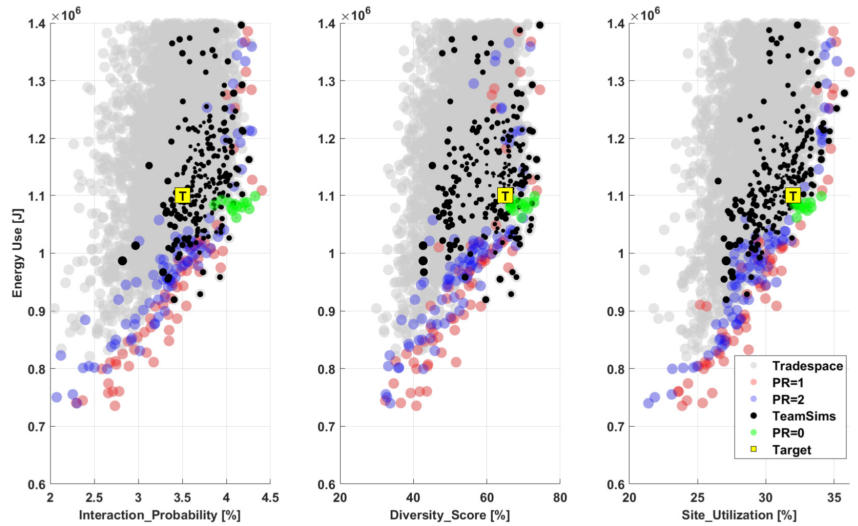


Figure 5-2: Visualizing Pareto Rankings in Star City - A multi-dimensional approach. Gray dots represent dominated solutions while the Pareto front is indicated at the bottom right. Red and Blue dots correspond to configurations with Pareto Rank 1 and 2. *Note 1*: PR1 would be the Pareto Line in a regular 2d space, but since the results are evaluated with multiple metrics, any Pareto Front does not form a line in 2d. This is why the blue and red dots mix in each subplot. *Note 2*: The black points represent simulations from the teams that did not meet the game requirements. The bigger dots are further away from meeting all requirements. *Note 3*: Green dots are simulations that meet all requirements.

"how good" their attempted Star City configurations are. Intuitively, this can be done by judging *how close* their solutions are to the Pareto front for every relevant tradespace. This process can be formally achieved using the Pareto Rank concept. In a nutshell, the solutions in the current Pareto Front are ranked 1; these solutions are removed (made infeasible), and the solutions in the new Pareto Front are ranked 2. The process is continued by identifying solutions in the front $k + 1$ after removing solutions in rank k [Goldberg, 1989]. By applying this process [4] to the Star City example tradespace, over 30 Pareto Ranks can be computed. Figure 5-2, on the right side, shows the first ten sets of Pareto Ranks.

A point must be made about the difficulty of visualizing a multidimensional optimization problem using 2d tradespaces. There are other ways to represent the multidimensional space (See Appendix B-3). In order to provide a notion of distance

[4] Alan de Freitas (2022). Pareto Fronts according to dominance relation. This function is available in the following link: https://www.mathworks.com/matlabcentral/fileexchange/37080-pareto-fronts-according-to-dominance-relation. MATLAB Central File Exchange. Retrieved July 13, 2022.

in 2d space, Figure 5-2 makes use of Euclidean distance (see Equation 5.4). However, an additional step has to be included: a rank lower than 1 for configurations that manage to meet requirements:

for i = 1:4 **do**
  if $x_{Si} meets y_{Ti}$ **then**
    $PR = 0$
  else
    $PR = [1, ..., 15]$
  end if
end for

Thus, teams that manage to meet requirements are also compensated for it, which regular Pareto Ranks would not acknowledge.

$$x_G = \sqrt{x_E^2 + x_I^2 + x_D^2 + x_S^2} \tag{5.4}$$

where x_G is the distance of the point to the required goal, x_E, x_I, x_D and x_S stand for the values of Energy, Interaction Probability, Diversity Score and Site Utilization.

Including a team's simulation run's results in the global tradespace makes it possible to get an impression of how much of it was explored. Figure 5-3 shows the location of all the results from a single team. The turquoise-to-violet color palette denotes the team's progression over time. For example, this team's earlier results are concentrated at the bottom-center of the tradespace, while their final results are placed nearby the bottom-right. This movement suggests that over time the team did get closer to a better location of the Pareto Front. Even though a visual analysis is not enough to conclude, it is a good result by the team and serves as a sanity check to contrast the Pareto Ranking.

All the pieces to estimate how well a team has performed are available at this point. The steps to obtain generalized performance metrics for all teams [5] are; (1) Group the results of the sampled Star City tradespace together with the results from

[5] Even though this process is followed for every team, the rest of the current Subsubsection only shows the results for one of them.

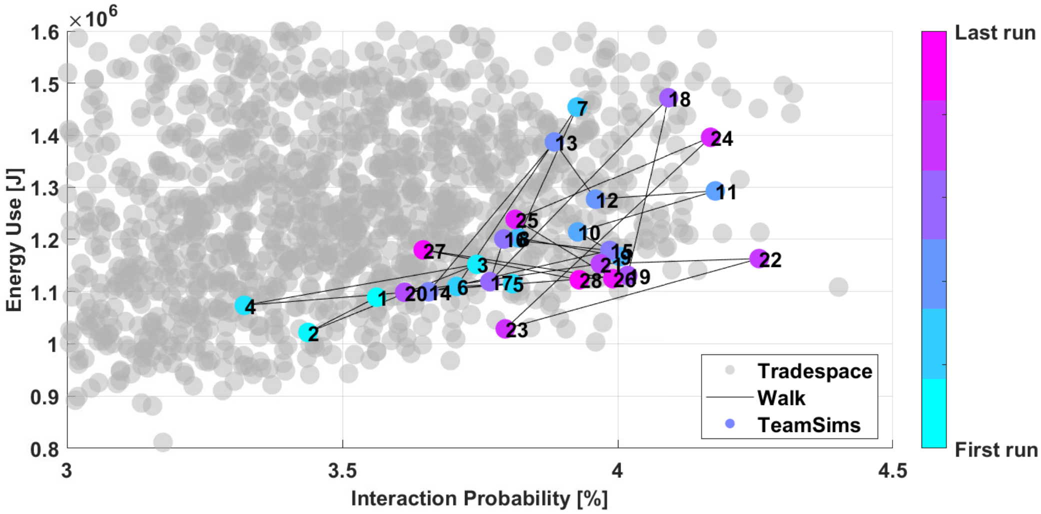


Figure 5-3: Single-team tradespace walk. It is worth mentioning a caveat of the current version of the time-shift visualization: the color distribution in the plot is equidistant among the points computed by the team, while the number in each point is its place on the team's simulation sequence. The actual time intervals from the team are not all equal. *Note 1*: The simulation count data has been normalized to account for the fact that there were 7 Control teams vs. 6 Treatment teams.

all the experiment teams, (2) Compute the Pareto Rank for all the points in the set and all four metrics [6], (3) Assign the Pareto Rank values to each team's simulations.

The Pareto Ranks for a team's simulations give a good impression of the local performance of the team using a global context. Figure 5-4 shows the same two metrics [7] for a single team's result set throughout the workshop. Briefly, it is clear that from all the simulations the team ran, only two were located in Rank 1 of the **global** Star City tradespace. Consequently, the Pareto Ranking for the rest of the team's simulations can be assessed and the team's average computed.

If there is interest to take a close look at the components of the *Total Changes* —orange line— signal in Figure 5-4, an example signal breakdown is provided in Appendix B-1.

Control vs Treatment Groups

This section's point of view is:

[6]Energy Use, Interaction Probability, Average Interaction Probability, and Average Diversity Score.

[7]Energy Use and Interaction Probability.

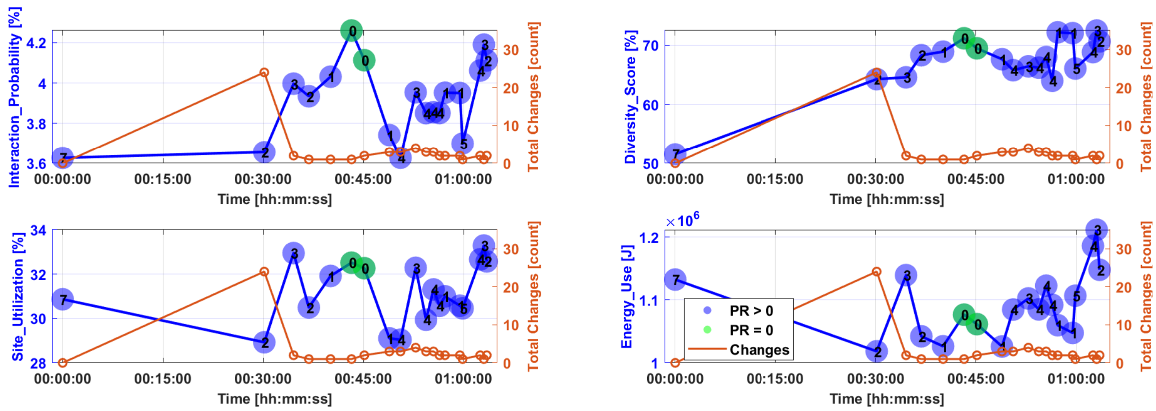


Figure 5-4: Timeseries for four FoMs and Pareto Ranks. Each one of this team's simulation runs has its corresponding PR. *Note 1*: Time in the *x-axis* has been modified to indicate only the duration of the Team Component of the experiment. *Note 2*: The orange line shows the number of changes a team performs from one simulation run to the next. Its maximum possible value is 35 — corresponding to the limits set by the SW user interface.

$\mathcal{K} \longleftrightarrow \mathcal{T}$ - Group to Group

Now that all teams' data has been quantified, it is possible to create an overview of the Pareto Ranking distribution for both groups. As mentioned before, each simulation run by a team is assigned a Pareto Rank value computed relative to the full tradespace — samples Star City plus other teams' data—. A simple visual inspection can give a first impression of each group's behavior, and since the results are discrete, a bar plot can be constructed. Figure 5-5 presents the distribution of Pareto Ranking values for Control and Treatment groups. The Control group distribution skews towards lower values than the Treatment set, indicating better performance. However, in this aggregated plot, it is impossible to see the performance on a team-by-team basis.

Figure 5-6 separates the Pareto Ranking metric from the Total Number of simulations. In this case, the number of simulations with Pareto Ranking = 1 is taken for every team on the y-axis. Simultaneously, the x-axis shows the total count of simulations [Manandhar, 2020]. This plot would indicate a better performance whenever a team achieves multiple PR=1 results, and if the team does so while carrying out many iterations, that presumes consistency of results.

The interpretation of this plot is more nuanced than a PR distribution. To ease

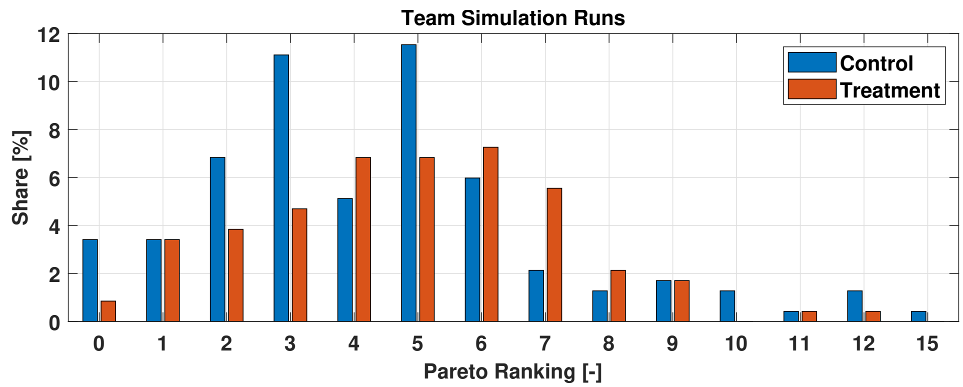


Figure 5-5: Distribution of Pareto Ranking by group type. *Note 1*: No teams had simulation runs with Pareto Rankings of 13 and 14.

the analysis of the results, the plot includes iso-lines [8] that can be used to compare teams concerning their distance from the utopia point [9]. As a cohort, the Control Group has more teams (T1, T5, T9) closer to Utopia, in contrast to the Treatment Group with a single team (T2), using the 7th iso-line as the cut-off.

There is a single following aspect to cover to characterize teams' performance: the longitudinal change, where pre- and post-treatment sections are appreciated. This information is not visible from the previous $\kappa \iff \tau$ point of view.

Pre- and Post-Treatment

This section's point of view is:

$$\kappa_{pre-post} \longleftrightarrow \tau_{pre-post} \text{ - Longitudinal Group to Group}$$

Following the same analysis and process described so far for all teams but now doing a split between pre- and post-treatment for the time series, I can get the breakdown presented in Table 5.1. For ease, the longitudinal evaluation for the κ and τ groups is done using the single split. One of the first compelling things is that most teams had a slow start regarding number simulation runs.

In this experiment, an assumption is made that a team with an improvement in average Pareto Rank — between pre- and post. Consequently, a decrease in PR

[8] A simple matrix multiplication was used to estimate the value of the distance from the utopia point.

[9] If the sample of points were larger than shown in the figure, a Pareto Front or computing new local Pareto Rankings of its own would be valuable for such a plot. However, since so few points exist, iso-lines make it easier to visualize.

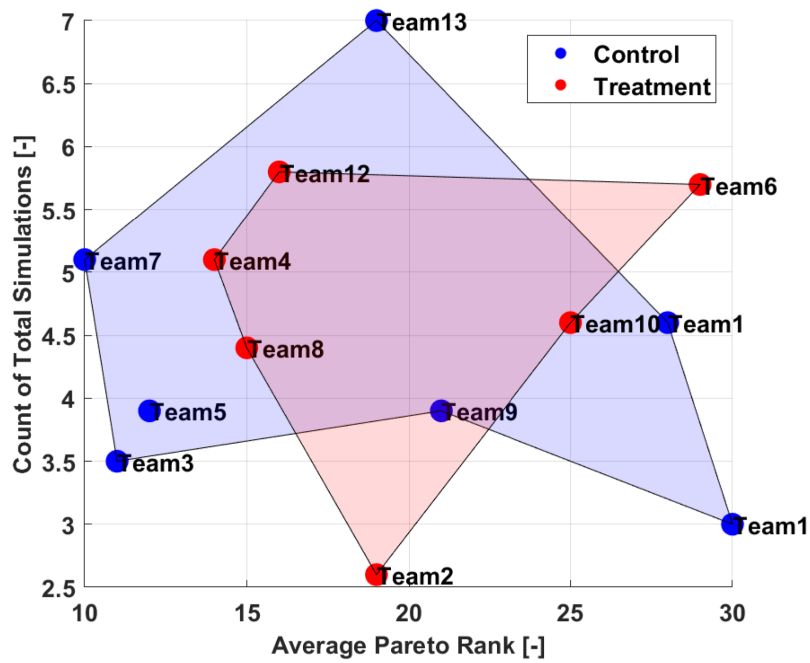


Figure 5-6: Total Simulations vs. Average Pareto Rank, by group. *Note 1*: Two teams show a Pareto Ranking = 0. Such value means that, compared to the global tradespace, those teams never achieved a simulation with Rank 1. *Note 2*: This plot shows only a subset of the data from Figure 5-5; that corresponding to Pareto Ranking = 1 — the two left-most bars.

indicates the worsening performance of a team.

Using the data from Table 5.1, I can create a visual overview that is more intuitive. Figure 5-7 shows the same information and insight. The vertical axis represents the average PR value pre-treatment, while the horizontal value corresponds to post-treatment. A team on top of the diagonal line would indicate that *on average* the PR performance did not change between the two stages. Consequently, the further orthogonal distance away from the diagonal represents a larger change between the two stages. The last two elements to describe are the vertical and horizontal lines, computed using the median of PR averages from the teams.

The final step is computing how many changes a team performed from one simulation run to the next. The Orgamachi User Interface, as used by the participants during the experiment, shows a total of 35 decisions relevant to the model; (1) 25 correspond to the population living-quarters allocation, (2) five to the transportation model per level, and (3) five to the communal space activity per level. The change

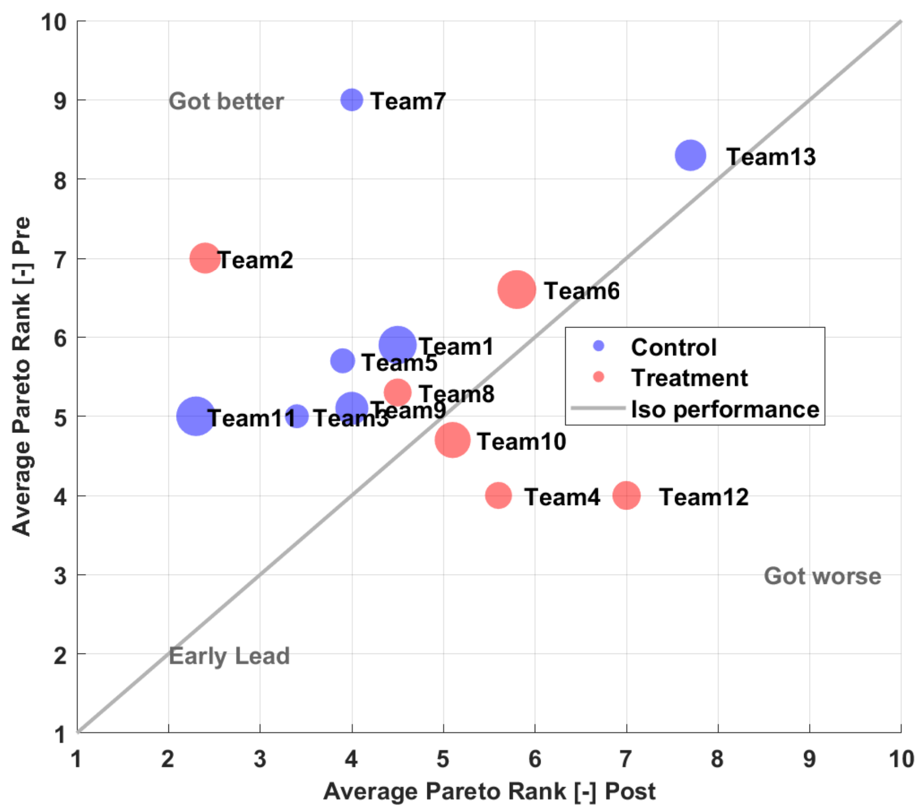


Figure 5-7: Pre- and Post-treatment Performance Distribution. *Note 1*: The circle's diameter indicates the number of simulation runs a given team performed. A bigger circle means that the team did more simulations. *Note 2*: The diagonal line would indicate that a team's performance averages from pre- and post-treatment are the same. The orthogonal distance between any team and the diagonal indicates a larger difference in that team's performance between pre- and post-.

				Number of Simulations		Mean Pareto Rank [-]		Standard Deviation Pareto Rank [-]	
Team Type	TeamID	Total Simulations	Mean Pareto Rank [-]	Pre	Post	Pre	Post	Pre	Post
Control	Team1	28	4.9	9	19	5.9	4.5	1.8	1.9
Control	Team3	11	3.7	2	9	5	3.4	1.4	4
Control	Team5	12	4.3	3	9	5.7	3.9	3.8	2.6
Control	Team7	10	5.5	3	7	9	4	8	1.4
Control	Team9	21	4.4	7	14	5.1	4	2	3.7
Control	Team11	30	3.3	11	19	5	2.3	3.5	1.6
Control'	Team13	19	7.8	3	16	8.3	7.7	2.1	3.1
Treatment'	Team2	19	2.7	1	18	7	2.4	0	1.5
Treatment'	Team4	14	5.5	1	13	4	5.6	0	2.4
Treatment'	Team6	29	6	8	21	6.6	5.8	3.2	1.8
Treatment'	Team8	15	4.7	3	12	5.3	4.5	2.1	2.9
Treatment'	Team10	25	5	6	19	4.7	5.1	2.1	1.8
Treatment'	Team12	16	6.6	2	14	4	7	1.4	3.8

Table 5.1: Pre and Post Metrics - Team Performance. *Note 1*: This table follows the same structure described in Figure 5-1. *Note 2*: The standard deviation of PR can be used as a better indicator of a team's accuracy at keeping system/problem performance. This metric is similar to the concept of "amount of PR=1 in the experiment" that was used in Figure 5-6.

metric used in this thesis analysis only counts the number of changes participants commit from one simulation run to the next. E.g., if the team changes every level's transportation mode and leaves the other two areas the same, the change metric would be 5. The maximum possible is 35, representing every decision option different from the previous simulation run. The minimum possible is 0, representing the team re-running the same site configuration.

5.1.2 Estimating Team Emotion

As mentioned in **Chapter 2**, emotion plays a role in the sensemaking process. In this experiment, each team's emotional journey is characterized using facial recognition software [10]. The software can generate two basic types of data; (1) *emotionality score*, ranging from -1 to 1, where negative values are associated with negative emotions and

[10]As mentioned already in Subsubsection 4.3.1 (see page 57), Moody can recognize seven emotions [11]: neutral, surprised, happy, fearful, disgusted, angry, and sad. The last three can be used given their association with a negative feeling.

positive values to positive emotions, and (2) breakdown of the emotions categories over time.

Metric definition

The emotion classifier predicts the probability of seven emotions for every timestamp. It provides a breakdown of probabilities for each emotion that indicates which one is the most likely to be occurring. In order to organize the progression of emotions over time, it is necessary to aggregate them into phases (see Definition 23), e.g., if an array of most likely —classified— emotions for six consecutive timestamps, it would include three emotion phases with different duration.

Definition 23 (Emotion Phase)**.** The sequence of multiple timestamps where the *Moody* classifier recognizes a singular emotion having the highest consecutive probability.

Figure 5-8 shows the emotion score data from one of the teams. Standard statistics such as mean, median, and standard deviation can be computed for each team and the pre- and post-treatment sections of the experiment. It is worth mentioning that the sampling rate from Moody is 0.5 [hz], which still creates a rather noisy signal. A Gaussian-weighted moving average filter with a window=50 is used to smooth the data. The rationale for the window size is that in a period of less than 25 [seconds], it would be unreasonable for a team-level emotion to change drastically.

Figure 5-9 shows a graphical representation of the emotion categories breakdown. Both subplots represent the same information. Although the left sub-figure is a more common representation, it is not easy to compare among multiple teams. The stacked bar representation on the right sub-figure is used instead.

In contrast to the emotion score, this data provides more granular information about specific emotions expressed by the team. In this regard, it is essential to highlight that a direct team-to-team comparison of emotion content, using either one of the variables presented so far, is not a good idea. Exploratory analysis shows teams' distinctive "levels" of emotion throughout the experiment, where some were

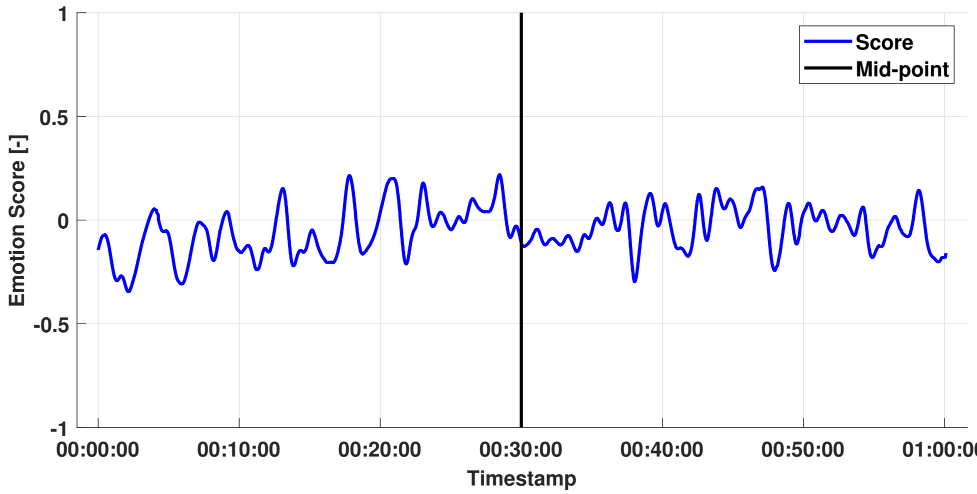


Figure 5-8: Emotion Score Timeline Example. The x-axis indicates the experiment time, and the y-axis is the output from Moody. The bold vertical line at the 30 [min] mark serves as a reference for the location of the *treatment* event during the experiment. However, this team is not necessarily in the treatment group.

markedly lower/higher than the rest without apparent reason.

Given that the level of aggregation for the work is only a pre- and post-treatment comparison, the following steps are followed on a team-by-team basis:

- For emotion score *numerical* data:

    1. Smooth the emotion score using a moving average filter.

    2. Compute *average* and *standard deviation* of the score for pre- and post-treatment sections.

- For emotion breakdown *categorical* data:

    1. Calculate the total emotion content share for the pre- and post-treatment sections: *aggregate emotion breakdown.*

    2. Calculate the delta in emotion content share between pre- and post- sections.

Overview of all teams

Figure 5-10 shows the change in emotion expressed by the team as reflected by the emotion score, as well as an explanation of how the plot should be read. The first

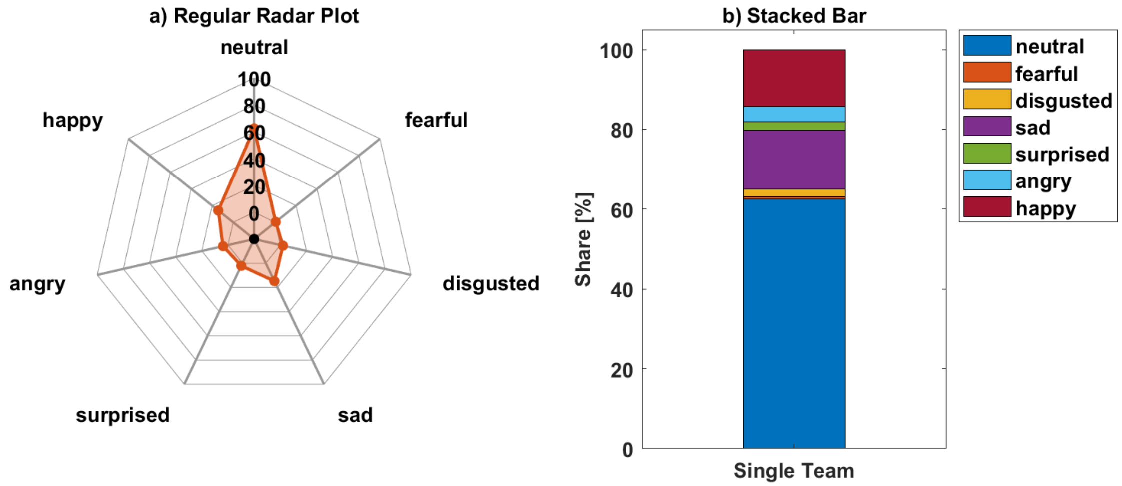


Figure 5-9: Emotion Categories breakdown - Single Team.

		Average Emotion Score [-]		Standard Deviation Emotion Score [-]	
Type	**TeamID**	**Pre**	**Post**	**Pre**	**Post**
Control	Team01	-0.06	-0.03	0.29	0.29
Control	Team03	-0.03	0.02	0.26	0.25
Control	Team05	-0.23	0.00	0.30	0.34
Control	Team07	-0.01	0.12	0.35	0.27
Control	Team09	-0.02	-0.01	0.27	0.25
Control	Team11	-0.05	-0.03	0.38	0.35
Control	Team13	-0.28	-0.18	0.28	0.36
Treatment	Team02	-0.16	-0.06	0.27	0.28
Treatment	Team04	-0.43	-0.23	0.29	0.26
Treatment	Team06	-0.05	0.07	0.30	0.27
Treatment	Team08	-0.13	-0.05	0.27	0.24
Treatment	Team10	-0.09	-0.06	0.26	0.21
Treatment	Team12	-0.20	-0.22	0.26	0.26

Table 5.2: Pre and Post Metrics - Emotion Score.

thing to notice in the plot is that teams with longer "connecting" lines between their pre-/post- sections show more change than the others, thus, a higher *emotion intensity* journey. The second thing to point out is that the "absolute" value of the emotion score, as was mentioned in the previous paragraph— has a high spread, even when only the team average values are shown. For example, Team 04's average pre-treatment emotion score is below -0.4 and goes up to approximately -0.22, which is a significant jump but still in the lower bracket compared to all the teams.

Meanwhile, Figure 5-11 shows emotion content for both experiment groups in the two sections. For example, Team 04 shows a noticeable difference between pre-

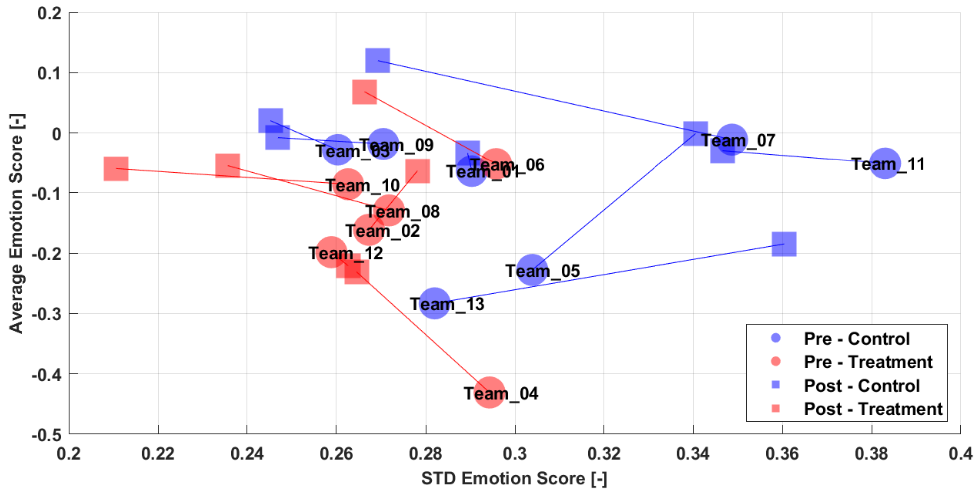


Figure 5-10: Pre- and Post- $\kappa \longleftrightarrow \tau$ by team Mean and Standard deviation. The x-axis shows the standard deviation, while the y-axis is the average. Control and treatment groups are distinguished by color (blue and red, correspondingly) and the pre-/post- sections by the shape of the data markers. The plot shows the team name closest to that team's pre- behavior. The length of the connecting lines denotes the level of change in score a team experienced between experiment sections.

and post- sections. It goes from ≈ 0.28 (aggregate of non-Neutral emotions) down to ≈ 0.21. Additionally, the "internal" share per emotion is also affected; *angry* is almost half its pre- size, and sad is $\approx 2/3$ of its pre- size. The reduction of certain emotions means that Neutral grew from ≈ 0.72 to ≈ 0.79.

Given that the teams ought to be compared based on their section-relative differences, the delta of team and emotion is shown in Table 5.3. In the top table, the rows are split with control (κ) groups at the top and treatment (τ) at the bottom. The color formatting in the top table is for all of its values, which helps to indicate which teams and emotions showed the largest change [12]. When the value is negative, the post- level is lower than the pre- value for that team/emotion combination. Positive values have complementary behavior, and the highest is indicated in green.

At this analysis stage, the statistical significance of the different levels is still not covered. However, even performing a basic comparison already shows interesting effects. For example, for both groups, the second section of the experiment showed a

[12]The *Neutral* emotion is left out of this visualization for the sake of simplicity. However, it is taken into account for the integrated analysis coming up in **Section 5.2** (page 91). It is also less informative given how the Moody algorithm computes it.

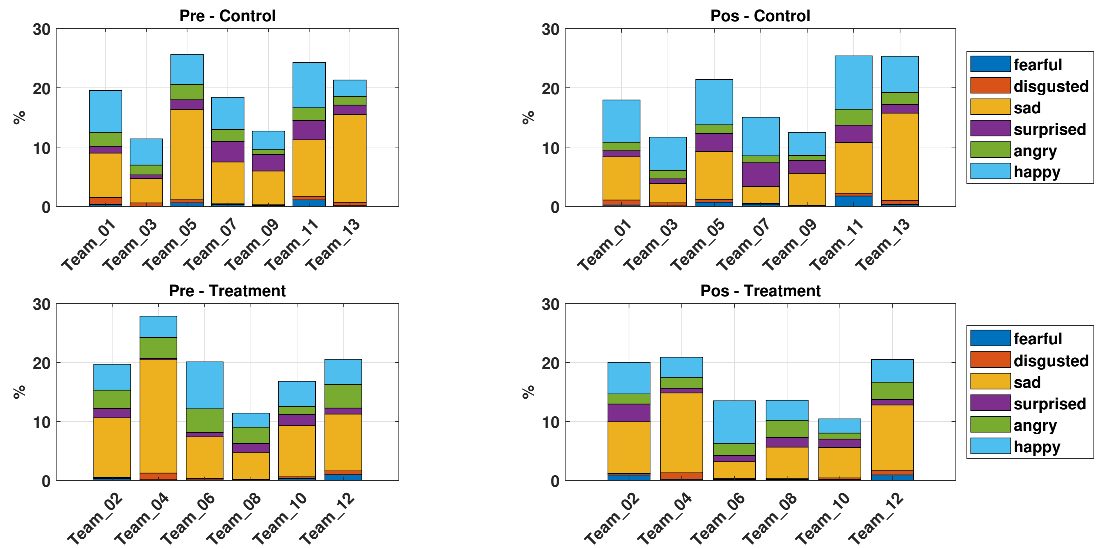


Figure 5-11: Pre- and Post- $\kappa \longleftrightarrow \tau$ by team Emotion Categories. The top two plots correspond to the κ group, while the bottom two plots are the τ group. Left plots are pre- and right plots post- sections. Each plot shows its corresponding team's aggregated emotion breakdown, which should be compared only with that team's complementary section. The y-axis is a unitary share (adding up to 1) of the total emotion category content. *Note 1*: In this plot, the "Neutral" emotion is left out for visual simplicity. *Note 2*: The colors in this plot do not match those of Figure 5-9, but the underlying data is the same.

decrease in the *Sad* category. However, only the Control Group shows an increase of *Happy*.

Individual Measurements from Smartwatches

Aside from the face recognition of emotions, participants also wore smartwatches to capture high-frequency physiological data; (1) heartbeat and (2) movement. Both data streams are captured as time series, which makes them good candidates for measuring the correlation of bodily signals within the team. Figure 5-12 shows an example of a team's physiological signals during the workshop.

For starters, the mean and variance for every participant are computed, and then the average over the team. As in previous sections, the time series are split into pre- and post- treatment. These would then grant the changes in magnitude and their variation.

In order to calculate team correlation, a linear interpolation of the heart rate data to 1 [sec] intervals is computed, together with its Pearson correlation. As an

		Delta from Pre- to Post-Emotion Breakdown [full percentage points]						
Type	**Team**	**Neutral**	**Fearful**	**Disgusted**	**Sad**	**Surprised**	**Angry**	**Happy**
Control	Team_01	1.8	-0.1	-0.3	-0.2	0.0	-0.9	0.0
Control	Team_03	4.9	0.0	0.0	-0.9	0.2	-0.2	1.2
Control	Team_05	-4.0	0.1	-0.1	-7.1	1.4	-1.1	2.6
Control	Team_07	3.6	0.1	0.0	-4.2	0.5	-0.8	1.1
Control	Team_09	6.9	0.0	-0.1	-0.3	-0.7	0.0	0.8
Control	Team_11	2.2	0.7	-0.1	-1.1	-0.3	0.5	1.3
Control	Team_13	6.2	0.2	0.1	-0.1	-0.1	0.5	3.4
Treatment	Team_02	3.1	0.6	0.1	-1.3	1.4	-1.4	1.0
Treatment	Team_04	13.7	0.1	-0.1	-5.6	0.5	-1.8	-0.1
Treatment	Team_06	0.7	0.1	0.0	-4.3	0.4	-2.1	-0.7
Treatment	Team_08	5.8	0.1	0.1	0.7	0.2	0.1	1.1
Treatment	Team_10	-0.2	-0.1	-0.1	-3.5	-0.5	-0.4	-1.8
Treatment	Team_12	-0.8	0.0	0.1	1.5	-0.1	-1.1	-0.4

Average	**Neutral**	**Fearful**	**Disgusted**	**Sad**	**Surprised**	**Angry**	**Happy**
Control	3.1	0.1	-0.1	-2.0	0.1	-0.3	1.5
Treatment	3.7	0.1	0.0	-2.1	0.3	-1.1	-0.2

Table 5.3: Pre- and Post- $\kappa \longleftrightarrow \tau$ by Emotion Categories. *Top table*: Shows the delta of emotion share, computed as $\delta = \kappa_{post} - \kappa_{pre}$, thus, having percentage points as units. It uses the data from Figure 5-11. The color distribution is based on the delta values for **all teams and all emotions**. *Bottom table*: presents the average **per team type** for each emotion through all the teams of a single type. *Note 1*: Both tables have red indicating a *decreased* post- level and green an *increased* level.

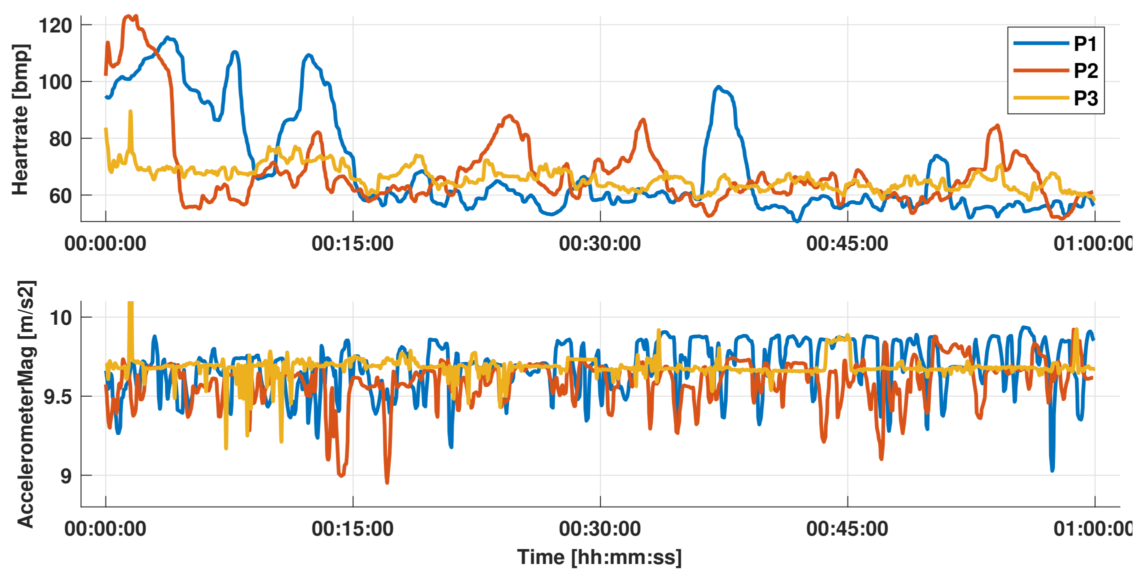


Figure 5-12: Team's physiological signals captured by smartwatches.

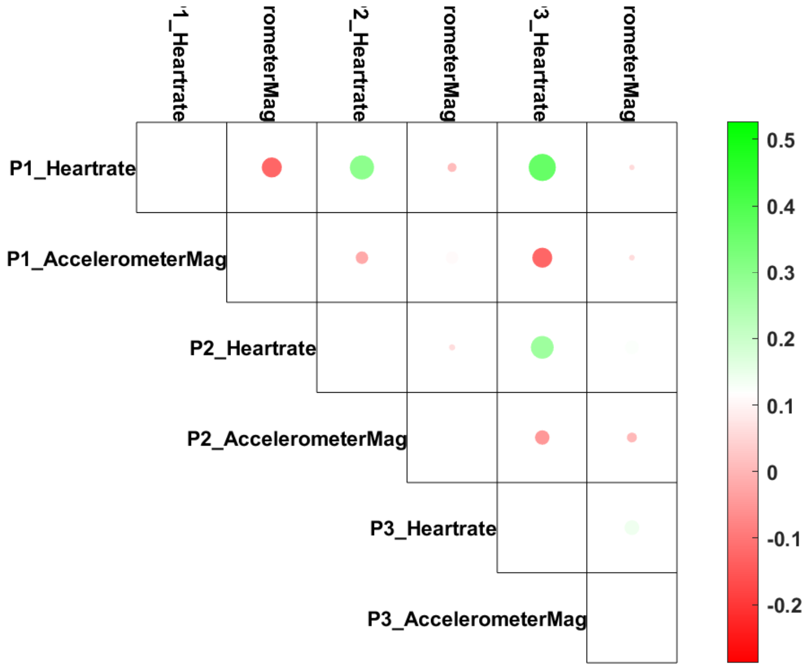


Figure 5-13: Correlation Matrix for Physiological Signals.

example, a correlation table for a single team is shown in Figure 5-13 as an example, indicating the team's level of entanglement [Gloor et al., 2022]. In the case of this specific team, the *heartrate* signal among the team members shows a high positive correlation compared to the accelerometer signal.

Finally, just as with previous sections, Table 5.4 provides the statistics pre- and post- for every team.

		Heart Rate Mean		Heart Rate STD		Acceleration Mean		Acceleration STD	
Type	**TeamID**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**
Control	Team01	55.4	51.9	4.8	2.2	9.7	9.7	0.2	0.3
Control	Team03	67.7	62.1	4.3	3.2	9.7	9.6	0.3	0.3
Control	Team05	64.0	66.2	9.6	5.9	9.4	9.4	0.5	0.6
Control	Team07	73.7	73.7	5.2	10.7	9.7	9.7	0.2	0.2
Control	Team09	65.8	62.8	4.4	5.5	9.5	9.5	0.4	0.4
Control	Team11	68.6	68.5	5.6	5.9	9.7	9.7	0.3	0.3
Control	Team13	73.1	70.0	6.0	7.5	9.4	9.4	0.5	0.7
Treatment	Team02	64.0	63.0	4.6	5.4	9.5	9.5	0.5	0.5
Treatment	Team04	72.6	62.7	15.5	7.2	9.5	9.6	0.5	0.5
Treatment	Team06	67.4	67.7	4.3	6.5	9.3	9.3	0.7	1.1
Treatment	Team08	71.3	64.7	9.3	10.6	9.5	9.5	0.5	0.4
Treatment	Team10	79.2	76.1	10.3	8.8	9.5	9.5	0.4	0.5
Treatment	Team12	72.9	69.9	7.6	6.0	9.6	9.5	0.5	0.5

Table 5.4: Pre and Post Metrics - Physiological Signals.

5.1.3 Estimating Team Conversation Pattern

Communication patterns can be visualized as a network graph, where a node represents each participant, and their contribution to the team discussion —relative to other team members— is depicted by the size of the node [Peng et al., 2022a]. In this study, a few metrics are computed to graphically ([Kim et al., 2012]) depict the dynamic behavior of a team's interaction based on their contribution to discussion equality and activeness of turn-taking. This subsection will introduce the definitions of the metrics used to characterize the communication patterns; then, it will show how those values are aggregated into graphs and time series for a couple of exemplary cases drawn from the teams that participated in the experiment. Finally, it presents the overview of communication behaviors in preparation for the integrated analysis —where communication is brought together with emotion and team performance in SubSection 5.2 (page 91).

Metric Definition

First, the *Individual Rate of Speaking* [Peng et al., 2022a] reflects how much time each individual is contributing to the overall communication for a given time window. Second, the *Pairwise Rate of Turn-taking* between two individuals. A response is considered to occur when the first individual begins an utterance within five seconds of the second individual speaking. The rate is the number of times this occurs in the time divided by the length of the period. This measure indicates which team members are interacting with which other team members and which people are involved in the same discussions *individual utterance rate* [Wiltshire at al., 2021].

Before continuing with the rest of the metrics, it is worth noting that the size of the time can have a confusing influence on the shape of the communication patterns ([Wiltshire at al., 2021], [Peng et al., 2022a]). To address this influence, it is possible to use the turn-taking *Degree of Change* (see Equation 5.5) to find the points that demarcate communication patterns differ from one another:

$$DC(t) = 1 - (x_{11}^{t-s} \; \ldots x_{ij}^{t-s} \; \ldots x_{nn}^{t-s}) \cdot (x_{11}^{t} \; \ldots x_{ij}^{t} \; \ldots x_{nn}^{t})(t \; \geq s), \tag{5.5}$$

where t is time, n is the number of team members, x_{ij}^{t} is the number of turn-taking between speaker i and j within t to $t+s$ minutes, s is the time span for comparing communication patterns.

Aside from information necessary to build the cliques over time, it is also possible to evaluate and aggregate the information from the clique to get more information about the team's dynamics. The evaluation can be done using participants' *Coefficients of Variance* (CV) from their conversations. The third metric used is the CV corresponding to the number of turn-taking between members, which can then be used to compute equality (see Equation 5.6) [Peng et al., 2022b]. A summary of these metrics is provided in Table 5.5.

$$Equality = 1 - \frac{CV_{Turn-taking}}{CV_{Turn-max}} \tag{5.6}$$

The final metric is the median speech length of every "speech event", For example, let's assume that during a certain period of time, the following sequence of [Participant n - time [s]] is followed: [**P1** - 3 [s]] → [**P2** → 4 [s]] → [**P1** - 5 [s]] → [**P4** - 3 [s]] → [**P1** - 2 [s]]. The median speech length is 3. This metric can also be computed for different lengths.

Metric	Code	Units	Description	Level of Aggregation
Pairwise Rate of Turn-taking	dialogue_num_sum	Count / min	Number of turn-taking per minute.	Node Pair
Individual Rate of Speaking	speech_len_sum	Seconds	Total speaking time per total time. Aggregate of all individual shares.	Team
Equality of Turn-taking	dia_cv	Dmls	Coefficient of Variance of the number of turn-taking between members.	Team
Median Speech Length	speech_len_med	Seconds	Median of speech length. Aggregate of all individual lengths.	Team

Table 5.5: Summary of Conversation Pattern Metrics.

Drawing communication pattern snapshots

Using each team member's *Individual Rate of Speaking* and the *Pairwise Rate of Turn-taking*, it is possible to construct communication cliques. The former metric defines the size of the circle for each one of the team members. It also determines the width of the connections between pairs of nodes.

Note: a team of three is as small as a team can get to have any semblance of an actual conversation pattern. Thus, some of the examples shown in the following paragraphs might come across as "intuitive," but that would certainly not be the case if the team had as few as two additional members.

Figure 5-14 shows four archetypal clique shapes present in the data captured from the teams. *Case A* serves as an introduction to the clique's graphic grammar: a team of three participants, each represented by a different color, shows different circle areas based on how much they spoke [13] during the evaluated time window. This case shows Participant 1 (**P1**) as having contributed the most and **P3** the least. Simultaneously, the connections between nodes show the proportion of turn-taking events per minute. *Case B* shows how the clique would look if a single participant **P2** seizes (almost-) all of the conversations. Importantly, this case also shows that in the case of two participants not addressing each other (**P1** and **P3**), the link disappears. *Case C* shows a more balanced team –compared to the previous two cases. Finally, *Case D* is how a clique would look in the plot if the team has less than six single turn-taking events in total. A non-conversation is still a pattern, but it ought to be differentiated

[13] Pairwise Rate of Turn-taking.

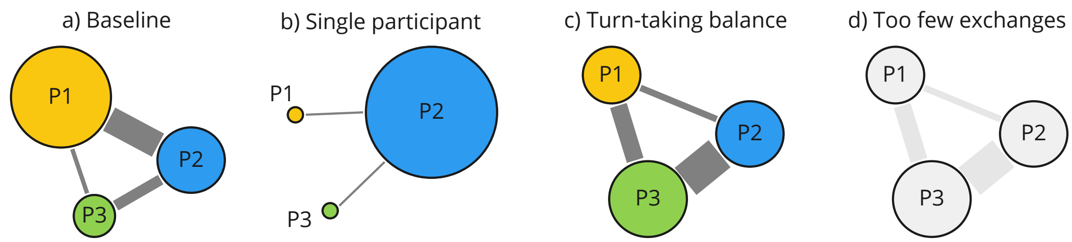


Figure 5-14: Graphic representation of communication cliques. *Note 1*: These figures do not correspond to any participant. They serve as mere examples.

from the rest because of its short duration, even though the values can be computed and represented graphically.

With this information at hand, all the information is available to integrate the communication journey for a team. Figure 5-15 shows one of the teams from the experiment. In contrast to other teams, this one shows a relatively simple sequence. Only three distinctive clique shapes are recognized, the third of which —starting around the 20 [min] mark— dominates the whole second part of the workshop. Looking at the first clique, one of the participants is not contributing verbally to teamwork, and the other two are engaged in active discussion. This same information can also be seen in the color bars at the bottom of the plot. The colors are the same as the cliques. The shape of the bars provides more granular information about which participant is speaking, when and for how long. The sharp drop in *equality of turn taking* —horizontal gray lines— is because, in the first clique, the green participant was contributing actively but shut down soon after.

Figure 5-16 shows the communication journey of Team 06. Based on this visualization technique and the numerical metrics, an evaluation of the performance-contributing factors becomes possible. In the same way as the previous Subsections, each team's hour-long workshop is split at the 30 [min] mark, where the Treatment Group received updated goals. The communication metrics are then evaluated for pre- and post- aggregates for all the teams.

Unfortunately, one of the audio recorders for a team failed at the start of their experiment and went unnoticed by the Facilitating Team until the end of the session. Therefore, one of the teams does not have any communication data.

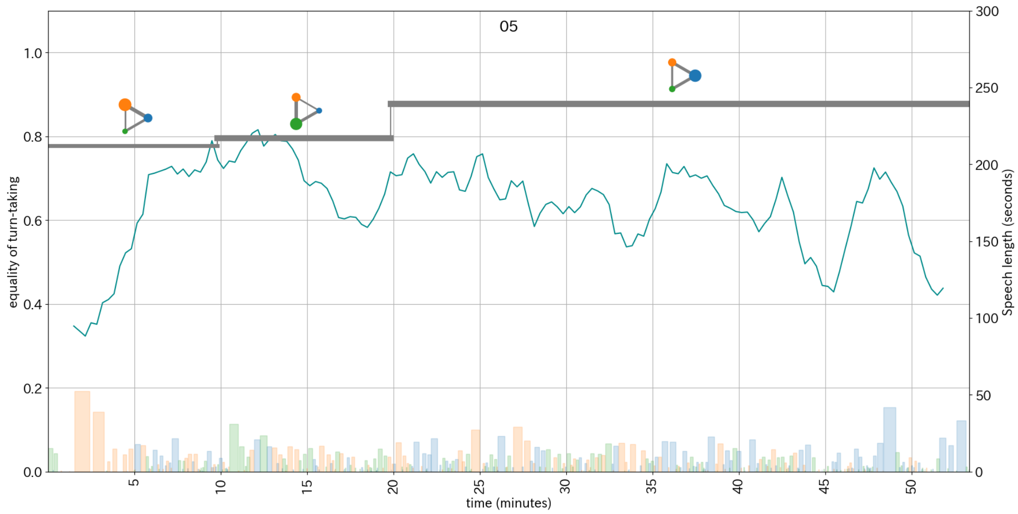


Figure 5-15: Conversation Pattern for Team 05. *Note 1*: The height of the gray line represents the equality of turn-taking. It is measured with respect to the left-side y-axis. *Note 2*: The color bars at the bottom represents which participants are speaking, for how long, and when. *Note 3*: The cliques represent the participants. The area corresponds to the time said member spoke. *Note 4*: The in-clique gray links' width is the number of turn-taking events between nodes. The aquamarine line —showing a more dynamic behavior— represents the team-level total speaking time every three minutes. *Note 5*: If the team is silent, the plot will show a space.

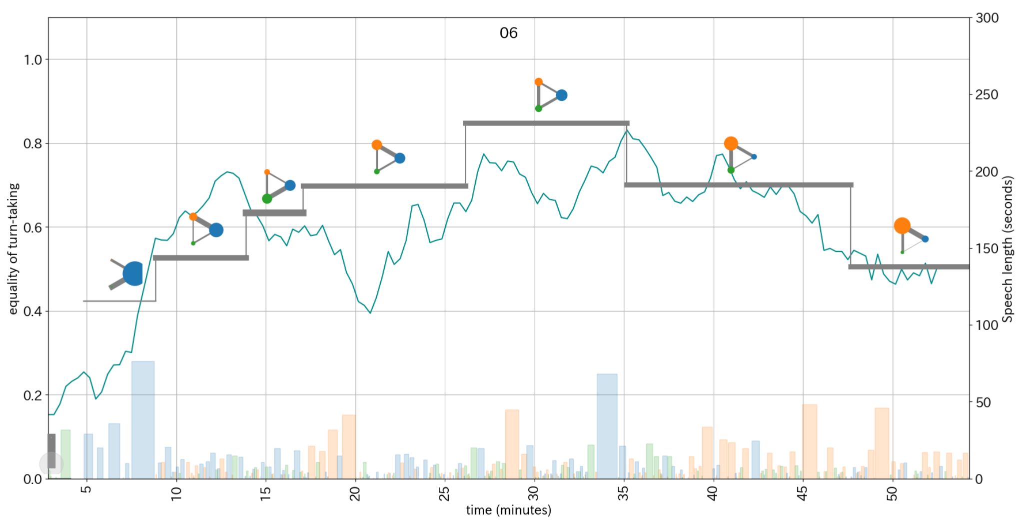


Figure 5-16: Conversation Pattern for Team 06. Note that this team has a far larger variety of clique shapes. It also has an initial period of silence and a "gray" clique —where the participants had less than six turn-taking events in a span of a few minutes— at the bottom-left corner of the plot.

		dialogue_num sum		speech_len sum		dia_cv		speech_len med	
Type	**TeamID**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**
Control	Team01	8.8	8.4	0.8	0.7	0.8	0.7	3400	3520
Control	Team05	6.7	7.8	0.6	0.6	0.9	0.9	3470	3225
Control	Team07	4.5	3.9	0.7	0.6	0.6	0.3	7620	5780
Control	Team09	5.6	6.4	0.6	0.6	0.8	0.6	4860	3670
Control	Team11	4.6	5.2	0.5	0.4	0.7	0.9	3420	3120
Control	Team13	6.7	8.8	0.7	0.7	0.8	0.8	3240	3420
Treatment	Team02	5.4	8.4	0.7	0.7	0.9	0.9	4750	3775
Treatment	Team04	2.8	6.4	0.5	0.7	0.6	0.7	5880	4340
Treatment	Team06	5.2	6.3	0.5	0.6	0.9	0.8	3975	4150
Treatment	Team08	7.3	8.7	0.7	0.8	0.8	0.8	3280	3710
Treatment	Team10	5.5	5.8	0.5	0.6	0.6	0.4	2940	4120
Treatment	Team12	5.9	7.6	0.5	0.6	0.8	0.8	3620	2940

Table 5.6: Pre and Post Metrics - Conversation Patterns.

Overview of all teams

With the team's workshop time split in half and aggregating the time series to a single metric, it is possible to construct Figure 5.6. Remember that one of the teams is missing from this table due to a technical problem with the measurement equipment.

Based on this data, the difference between Post- and Pre- can be calculated on a team-by-team basis. Table B-4 shows the comparable data for all the teams with the Performance and Emotion metrics.

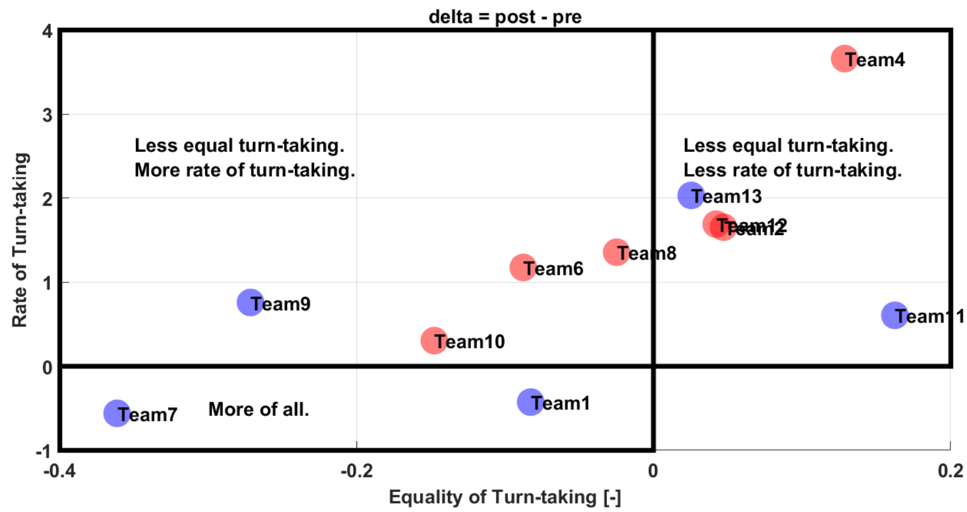


Figure 5-17: Turn taking equality patterns. This plot shows one of the possible visualizations indicate differences in communication patterns by the teams, specifically when comparing pre- and post- conversation metrics. The two metrics under analysis are; (1) equality of turn-taking in the x-axis and (2) rate of turn-taking in the y-axis. *Note 1*: The bottom-left corner teams show teams that increased both behaviors in the second half of the experiment. Meanwhile, the top-left corner shows teams with less equal turn-taking but also high rate of turn-taking. Both of these corners would indicate an increased interaction in the discussion. Although it is not possible to make an attribution as to *why* teams showed that behavior, the following section in the thesis will link it to teams performance.

5.2 Bringing it all together

This Section presents a comparison across the different data streams —team performance, emotion and conversation. It also explores the relation between team performance and the different behaviors shown by the teams, including the difference between Control and Treatment Groups. Before continuing, it is important to recall that the team performance, as measured by the Pareto Ranks, was assessed based on the *original* team goals —before the Treatment group was informed that their goals had become more stringent. Thus, the Treatment group is being evaluated equally as the Control.

All the data used as input for the analysis in this section can be found in Appendix B-4 and B-5. It is a collection of the summary tables presented at the end of each of the previous sections.

Lastly, the original research questions posed by this thesis are:

RQ1. How does a team's emotional journey during teamwork influence the

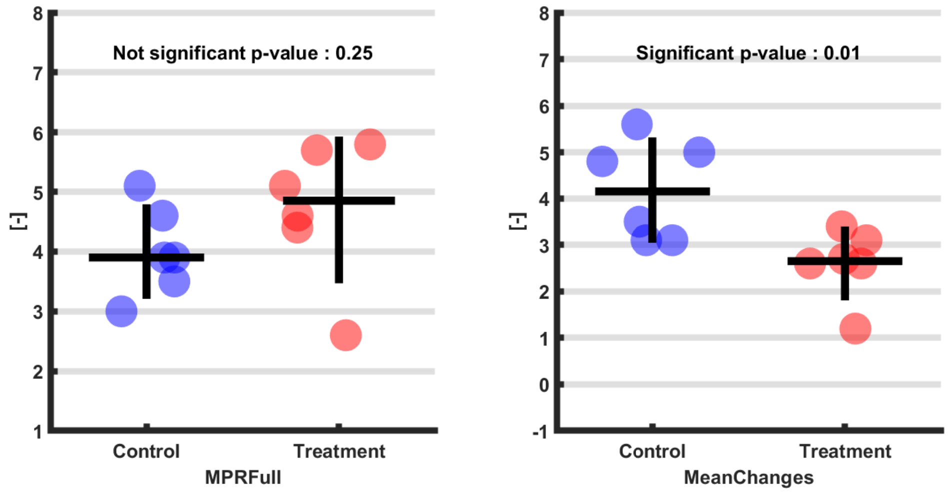


Figure 5-18: TestHypohtesis1A. *Note 1*: Both axes are computed for the full-hour experiment. *Note 2*: The right-side plot does not indicate team performance but rather a *team action*: how many design changes, on average, did the team make from one simulation run to the next.

team's performance?

RQ2. How do conversation patterns influence team performance?

Given the low number of teams that participated in this experiment, it is not possible to make any statistically significant claims that can shed light about the aforementioned research questions. However, it is my belief that the methods used to capture team performance and behavior can be reproduced with high n and followed in a similar fashion to strengthen this thesis' point of view.

The first important piece of information is to establish whether there is a difference in performance between Control and Treatment. Figure 5-18 presents the results of a t-test for the primary performance metric: (Total) Pareto Rank and one team behavior: the average number of design changes in each simulation run. The left plot, Performance Rank, does not show a statistically significant difference in performance. However, it does show that, as a group, the Control cohort has lower (better) results. Simultaneously, the right-side plot shows the first statistically significant behavior between the groups: the (average) amount of design changes each cohort performed between simulation runs. The Control group decidedly performed more changes compared to the Treatment.

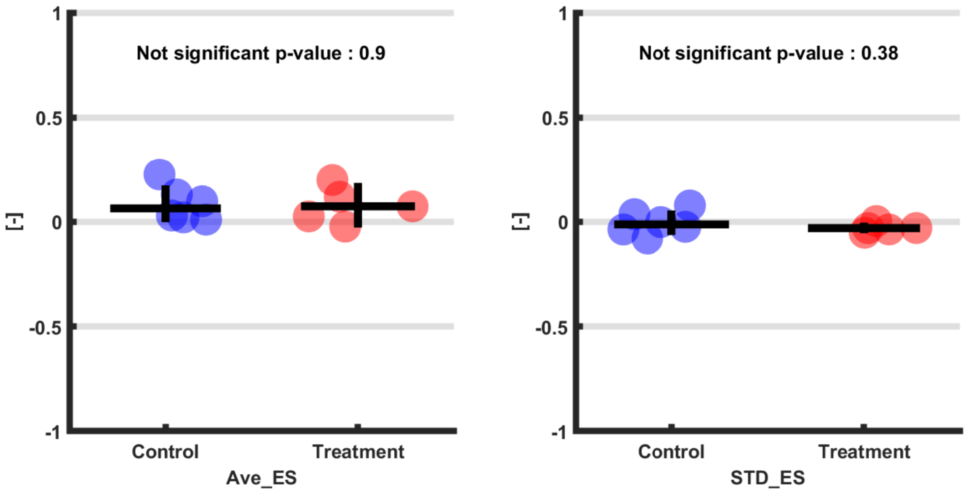


Figure 5-19: TestHypohtesis1A. *Note 1*: The y-axis for both plots is the difference of average values between sections pre- and post-. Thus, if the value is positive, it means that the post-level had a smaller value than the pre- section for a given team.

5.2.1 Testing the hypotheses

In order to give more context to the research questions, let us look at the research hypotheses. The purpose here is to establish whether there was any difference in the behavior of the teams depending on their group type (Control / Treatment). Here is Hypothesis 1:

Hypothesis (H1)**.** Teams that experience a (1) more **intense** emotional journey and (2) **moderate frustration** during the process of sensemaking have higher performance,

According to a t-test result, based on **H1**, the intensity of the emotional journey can be interpreted to be a difference in the emotional score captured during the experiment. Thus, a team with a more significant difference between the experiment sections would have had a more intense journey. Similarly, larger shorter-term variations, captured by the signal's variance, indicate a sense of intensity. Figure 5-19 shows no statistical difference between Control and Treatment groups.

The sensors in the experiment do not capture frustration by themselves. However, they capture levels of Happiness and how much participants show Anger. Figure 5-20 shows no statistical difference in the *level* of changes between pre- and post- sections

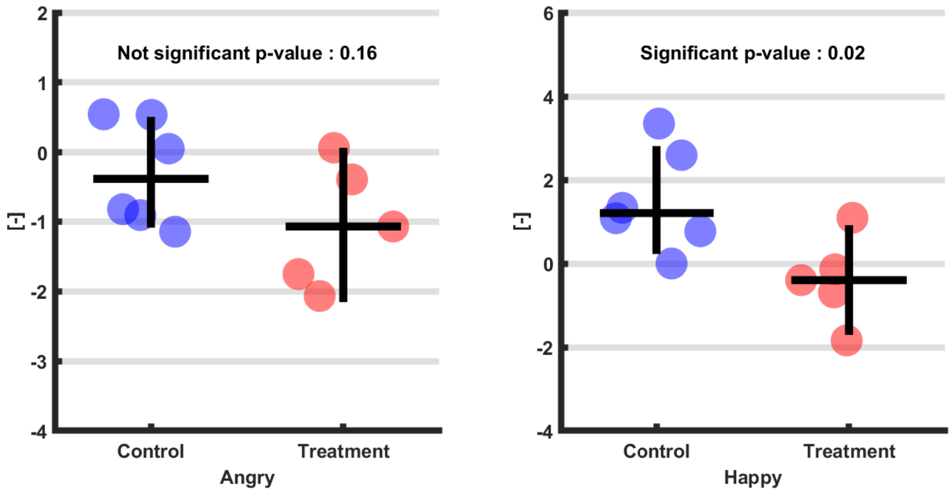


Figure 5-20: TestHypohtesis1B. *Note 1*: The y-axis for both plots is the difference of average values between sections pre- and post-. Thus, if the value is positive, the post-level had a smaller value than the pre- section for a given team. *Note 2*: The horizontal black line for each group indicates the group's median. *Note 3*: The vertical line for each group indicates the Confidence Interval at 95% confidence for each group.

between the Control and Treatment groups.

Meanwhile, the same metric of *change between sections* is applied to happiness on the right side of Figure 5-20 and a statistical difference is detected. Being in the Treatment Group harmed the happiness levels of participants post-treatment.

In brief, the only statistically significant factor is that Treatment teams showed a lower happiness level, compared to Control, in the post- section of the experiment —after they were informed that their requirements had become more stringent. No other emotional factors indicated significant differences.

Following up with the second hypothesis of the experiment:

Hypothesis (H2)**.** Teams that have an **equitable** and **stable** participation, among their members, during a work session have a higher performance.

The first check is to verify whether there was any behavioral difference between the groups. Equitable participation during the team discussions is captured by the **dia_cv** metric. The right side of Figure 5-21 shows no statistical difference between Control and Treatment.

Meanwhile, the stability of the participation of members can be drawn from the

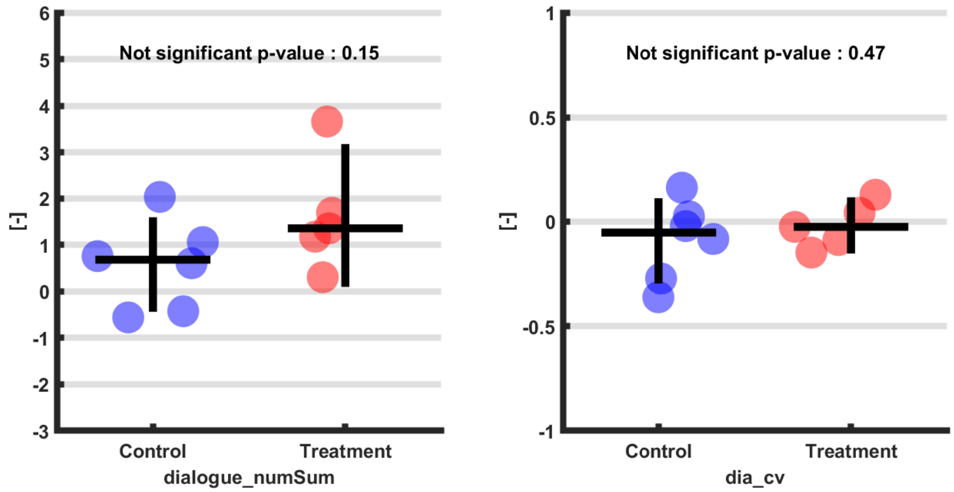


Figure 5-21: TestHypohtesis2A. *Note 1*: The y-axis for both plots is the difference of average values between sections pre- and post-. Thus, if the value is positive, the post-level had a smaller value than the pre- section for a given team. *Note 2*: The horizontal black line for each group indicates the group's median. *Note 3*: The vertical line for each group indicates the Confidence Interval at 95% confidence for each group.

total number of dialogues change. The left plot of Figure 5-21 shows no statistical difference between the groups. However, it does show a more considerable difference compared to the **dia_cv** metric.

The previous statistical tests suggest that the only clear behavioral difference in this experiment —directly linked to the original hypothesis— is how happy teams felt. At the same time, teams that perform multiple changes between simulations seem to benefit in terms of their final performance. With this indication, it is time to examine how the measured behaviors contribute to the final Pareto Rank.

Using a linear model, the data were fitted, one factor at a time, to the (Total) Pareto Rank. The results are shown in Figure 5.7. The table splits the contributing factors into the general categories of Emotion, Communication, and *Others*, where the latter shows factors related to physiological measures and; (1) how many simulations the teams performed in total, and (2) how many design changes were made between simulations.

Since it has already been established that most of the behaviors between Treatment and Control are not statistically different, these linear regression models use the data from all the teams as a single block. The table has cells marked in yellow per

Dependent Variable: (Total) **Mean Pareto Rank**

	Variable	p-value
Communication	speech_lenSum'	0.36
	speech_lenMed'	0.66
	dia_cv'	0.74
	dialogue_numSum'	0.92
Others	AccelerationSTD'	0.15
	MeanChanges'	0.19
	AccelerationMean'	0.26
	HeartRateMean'	0.47
	TotalSims'	0.65
	HeartRateSTD'	0.87

	Variable	p-value
Emotion	Fearful'	0.07
	Surprised'	0.39
	STD_ES'	0.42
	Disgusted'	0.42
	Neutral'	0.66
	Sad'	0.84
	Ave_ES'	0.87
	Angry'	0.90

Table 5.7: P-values for regression model of all measured factors vs. Mean Pareto Rank.

category with the lowest —relatively more significant— p-value.

First, in the *Communication* category, **speech_lenSum** —how much the teams spoke in a given time — is the one with the lowest p-value, although it is far from being statistically significant. Interestingly, the **dia_cv** metric —related to the equality of conversation— has a high p-value, which seems to contradict common wisdom about high-performing teams being the ones with equitable discussion. Finally, **dialogue_numSum**, the total number of turn-taking events in a time — has the lowest influence on team performance.

Second, the *Emotion* category, shows that the **Fearful** score has the lowest p-value. The interpretation for this metric indicates that teams with a higher Fear post-treatment achieved better performance. Figure 5-22 shows the relationship between these two variables.

Meanwhile, on the other side of the p-value spectrum for the Emotion category is *Angry*: the change in this emotion amount between the first and second half of the experiment showed no influence on team performance.

Third, the *Other* category shows that **Acceleration Standard Deviation** —captured by the teams' smartwatches— is the better p-value to describe Team Performance. A close second would be the **amount of design changes** the team performed in-between runs. Interestingly, the **Total number of simulations** a team performs

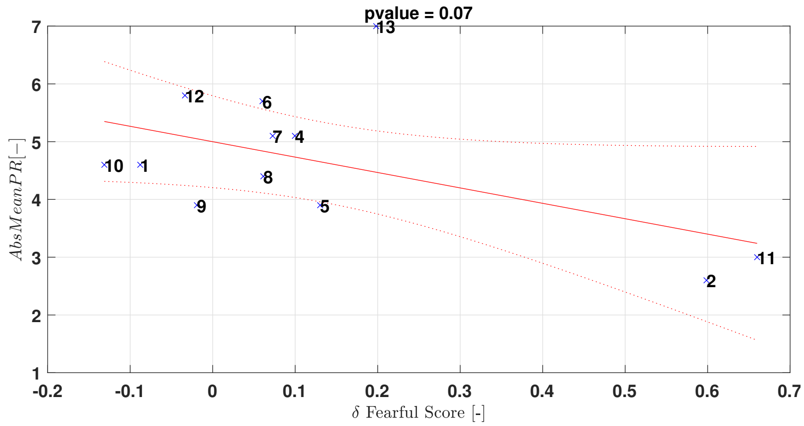


Figure 5-22: Linear Regression Model - Fear Score. The numbers next to the points in the plot are the team IDs.

has a statistically significant p-value score.

Based on these results, a key takeaway suggested is that a good team performance —within the boundaries of this experiment— can be obtained by running fewer simulations if the team makes numerous design changes after a lengthy initial discussion.

5.3 Summary

This chapter presented the approaches and tools used to estimate team-level performance and behavior. A modified Pareto-Ranking approach was used to assess the team's performance during the game. Emotion is measured using a facial-recognition algorithm that estimates average team emotion intensity and general categories of feelings. Lastly, voice recognition is used to compute and compare communication patterns between the experiment participants during the game. These scores and metrics are then used to assess the influence of emotion and communication on team-level performance.

Chapter 6

Discussion and Next Steps

> Just keep swimming.
>
> Dory, *Finding Nemo* (2003).

This Chapter closes the research done in this thesis. **Section 6.1** takes the results presented in the previous chapter and elaborates on what insights can be drawn from them, as well as the relevance of the proposed method to the science of teamwork. **Section 6.3** highlights the known limitations of this thesis, its caveats, and areas that can be strengthened in future work aimed at using/reproducing the methods and results used here. **Section 6.2** looks forward and lists a few options of exciting work to be done in sensemaking instrumentation at the meso-scale (team-level). Second-to-last, **Section 6.4** focuses on recommendations for future work based on lessons learned during the design and execution of this thesis research agenda. Finally, **Section 6.5** provides key takeaways from this chapter.

6.1 Discussion

One question remains with all the results already presented: *so what?* Let us start with the key findings from this research; (1) at the time scale of the experiment, it is possible to observe and measure specific team-level behaviors related to Weick's sensemaking properties, and (2) none of the behaviors measured showed a clear and statistically significant influence over a team's overall performance —as measured by the Average Pareto Rank.

One of Weick's properties, *sensemaking is about presumption and action*, can be observed during the experiment presented in the thesis. Multiple teams took a long time to discuss during the beginning of the workshop without engaging a lot with the simulator —even though they were not explicitly asked to do this, and there was no cost or penalty to running the simulations. One could argue that the team, as a unit, needs to "understand" the problem before engaging with it.

Another property for which it is possible to capture a pattern is the one related to the *social and systemic* nature of sensemaking. The interaction between the team and the simulator is captured by both the communication pattern followed by the team and the tradespace exploration [1] of the system; in addition to quantifying the number of design changes performed by each team. The communication patterns also show that specific individuals also take charge of the team conversation at certain times. Although the correlation between the role assigned to those participants and the types of changes they performed is not studied in this thesis, the data is available to look at those correlations.

One thing was very clear from the tradespace walk of the teams: it was not random. Every team was able to get closer to the multidimensional Pareto Border at some point and without the need for dozens of simulation runs. This has the potential to be attributed to the *noticing and bracketing* and *labeling* properties of sensemaking. The Mars' Star City simulator might be a black box for the participants, but the design options provided do give intuition to the teams by the mere fact of "making sense." For example, listening to the audio recording from the teams, many of them quickly and correctly— assume that one of the transportation options will require a lot more energy resources that are unlikely to bring them closer to the Pareto Border. This clue embedded in the language and semantics of the user interface allows the participants to make quick progress on which combinations of options could be

[1] The Mar's Star City user interface shows an empty tradespace at the beginning that becomes populated with the simulations performed by the team. It never presents an overview of the theoretical tradespace, as the intention is for the participants to *discover* its shape.

more valuable.

Compared to previous work from Global Teamwork Lab members, this research aimed to contribute to measuring the team dynamics and patterns and how they relate to the team's exploration of the system. In that regard, there was little space for unexpected results. Even though the treatment applied to one of the experiment groups seems to have been counterproductive to their performance.

6.2 Next steps

This thesis work required a narrow interpretation of sensemaking to become measurable in a short time and for a small group of people. A clear next step can be to carry out a cross-sectional study using the same game and method but with a different population. Ideally, it should still be comparable, with a limited amount of context changing. For example, the same game used in this research can be executed in an organization with a markedly different cultural pattern or with different communication dynamics and patterns [Maitlis, 2005]. The experiment could be a simple problem to characterize how different groups of people engage with the same type of problem. Even more so, real-time overviews of the teams could also be constructed as an educational tool —see Figure B-6 for an example.

In addition, the teamwork visualization techniques used here can also be valuable for pedagogical work [Paasivaara et al., 2014]. Furthermore, a longitudinal study in a group that repeats a similar process every year can also be interesting. The current version of the work presented in this thesis does not account for *long-term* sensemaking. Even though the instruments are not meant to be used on a scale over one to three hours, the principle could be the same. What would a sentiment or communication pattern look like on a month-scale for hundreds or thousands of people? New instruments and methods would be necessary for this.

6.2.1 Distributed Sensemaking (at Scale)

Big organizations engage in sensemaking differently when the patterns they try to detect and enact emerge, even more so when it is an unexpected emergence [Weick, 2012]. Instrumenting the sensemaking process with a focus on social communication and emotional aspects would require different instruments. Some of the main challenges are; (1) the protracted unfolding of events of interest [Weick, 2012], (2) organizational inefficiencies [Wolbers, 2021], and (3) the ambiguity of causality [Juarrero, 2000].

The use of additional material is a better fit for the different cognitive preferences of the team members [Stigliani & Ravasi, 2012]. Even in the experiment presented in this thesis, it was noticeable how teams used paper documents, ranging from sketches and notes to folding them in shapes to make a point to other teammates. The role of ethno-materiality could also play a role in how teams perform and interact. Further work can bring value to sensemaking, especially in long-term processes.

6.2.2 The importance of content

In larger scales of time and when information about the product/system is uncertain, the capacity of individuals to assess the accuracy and usefulness of their knowledge becomes paramount. I do not think there is an instrument capable of judging this in a natural environment, but it could be possible to set up experiments and check for *mature experts* to see whether it is possible to know how reliable they are [Tetlock, 2005]. A research opportunity here can be the evaluation of the argumentation pattern of individuals in a team using metrics such as the *conceptual/integrative complexity score* (CIC) [Baker-Brown et al., 1992] [2].

[2] Unidimensional score on a 1-7 scale. Scores of 1 indicate no evidence of either differentiation or integration. Scores of 3 indicate moderate or even high differentiation but no integration. Scores of 5 indicate moderate to high differentiation and moderate integration. Scores of 7 indicate high differentiation and high integration. Scores 2, 4, and 6 represent transitional levels in conceptual structure.
Some key features are relevant to mention about the CIC score:

1. The scoring system focuses on structure rather than content.
2. It is essential not to allow the coder's personal preferences or biases on an issue to influence

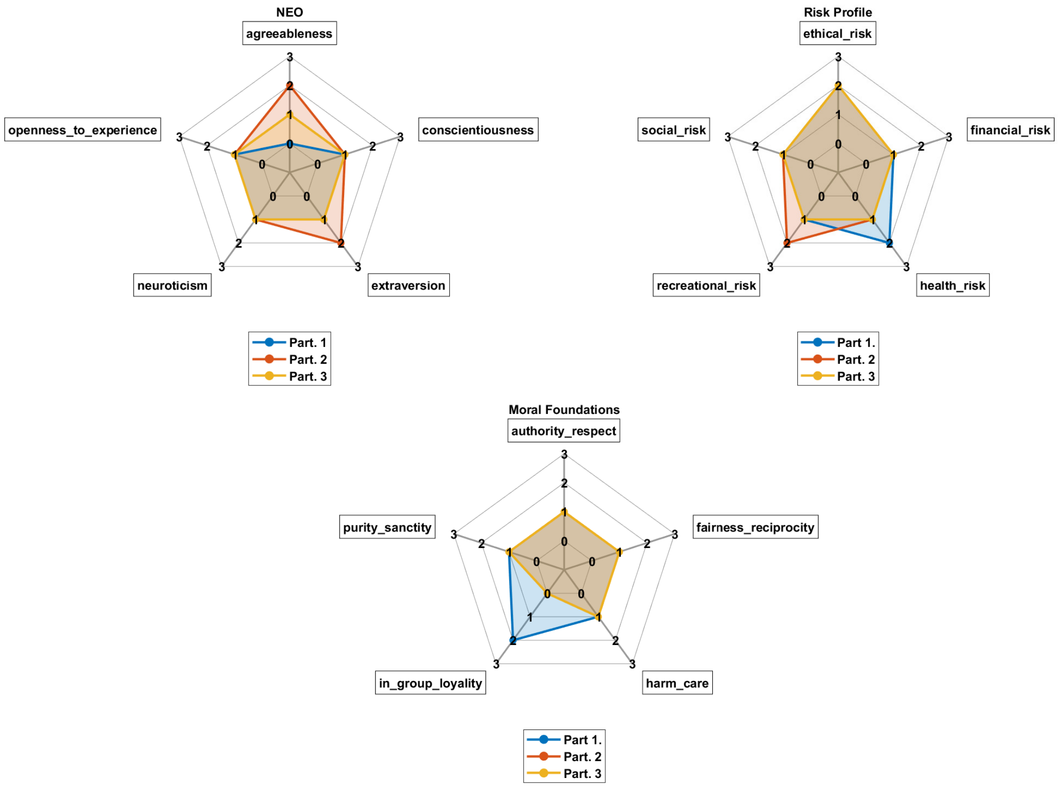


Figure 6-1: Sample individual personality assessment for a single team.

6.2.3 Individual and Team Personalities

Finally, the role of team personalities on how team-level dynamics end up playing is also possible to investigate. Even though the data from the individual members were captured for this experiment, the work of verifying and validating the process of deriving accurate personality assessments from the video capture proved to be lengthier than expected. Figure 6-1 shows a sample of what a single team's personality profiles could look like. A breakdown of three personality tests, (1) NEO, (2) Risk Profile, and (3) Moral Foundations, can be drawn for each participant and compared against communication patterns and also overall team performance.

---

the conceptual assessment of a statement.

3. The coder should not always assume that it is better to be more complex.

6.3 Limitations

There is a distinct lack of empirical work to take seriously the nature of "embodiment" in the workspace [Hindmarsh & Pilnick, 2007].

The metrics used in this work still rely on averaging relatively long portions of time, and they all discard the "content" of the conversation. Thus, making it infeasible to detect whether a team is a *silent genius*, whose contribution to the team's performance cannot be quantified by the amount of time spent in the conversation but by the impact of her/his insights.

6.4 Recommendations for future work

1. **Double-check the influence of randomized behavior in the simulator used by the teams**. In this experiment, even though the teams all started with the same configuration, the stochastic nature of the simulator caused the results of that simulation run to variate by multiple Pareto Ranks. This means that if teams run a simulation with a good configuration, but it is unfortunate to have a low performance, the signal to the teams can become murky, and the participants can take the wrong insights from experience.

2. **Execute multiple dry-runs**. An ordered and frequent dry run regime before the actual experiment is key to having a successful and scientifically valid—result. The dry run not only brings information to the surface about the measurement equipment and how the participants interpret the game instructions but also provides valuable experience to the Facilitator Team.

3. **Have a simple but strong treatment event**. One of the early decisions for this experiment was to minimize the number and type of interventions on the team. Teamwork experiments usually suffer from low participation, which makes it challenging to run meaningful statistics. In addition, research teams also try to test multiple treatments during the same experiment, causing a *fishing in a small barrel* challenge. It is easy to end up with too few teams to

observe anything. However, this experiment could have gone too far to the other side. There is no statistical evidence that the treatment used in this experiment impacted the team performance. So, it was kept simple but not strong enough.

6.5 Summary

This chapter presented some reflections on this thesis work's contributions, limitations, and potential research avenues. One thing is clear; there are various ways to continue bringing insights and creating knowledge not only about *how* to measure teamwork but also about the implications of the phenomena being measured. I am not quite sure of having a better way to summarize the challenges of this corner of science than what a previous Global Teamwork Lab contributor already said: **Teamwork research is hard** [McDonough, 2021].

Appendix A

Workshop Material

All right! We did not die today. I call that an unqualified success.

Fear, *Inside Out* (2015).

Architectural Decision	Option 1	Option 2	Option 3
1. Time Scale	*Minutes*	*Hours*	*Days*
2. Unit of Analysis (UoA)	*Individual*	*Team (3-5 members)*	*Multi-team*
3. UoA Geographic Separation	*Isolated* *Participants are essentially playing alone from many separate locations.*	*Clustered* *Participants gather in small groups at specific locations.*	*Co-located* *Participants gather all together at one location.*
4. Time Separation	*Asynchronous* *Participants are online at different times. Live interaction is limited to those online at any given time. Game play starts and stops locally at different times.*	*Staggered* *Participant play overlaps at specific times for live interaction; the remainder of the time play is essentially asynchronous.*	*Synchronous* *All participants are present and interacting at the same time. Game play starts and stops for all players simultaneously.*
5. Tools	*Ad Hoc Tools* *Applications not designed specifically for gaming but can be adopted for the purpose.*	*Dedicated Tools* *Applications designed and optimized for gaming.*	--

Figure A-1: Morphological matrix with experiment design options.

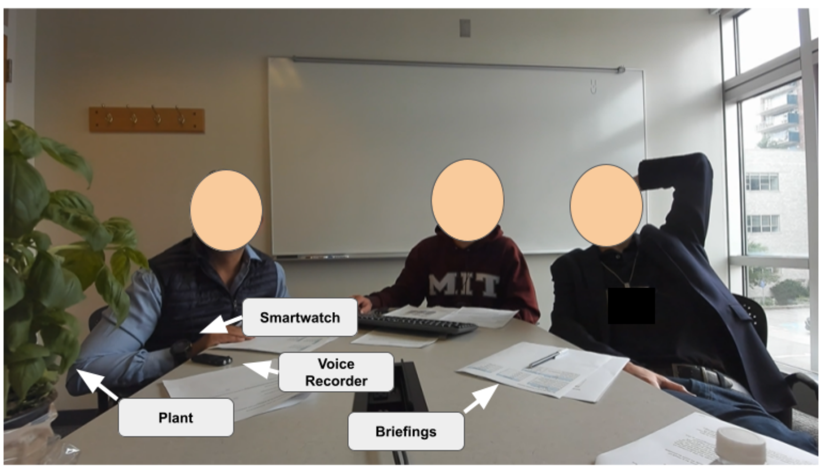


Figure A-2: Team working in a breakout room.

Mars Star City Design & Simulation Workshop
Team Briefing

Team Briefing

StarCity Mission

Advancing colony construction, in-site resource utilization, and additive manufacturing on Mars.

Team Level Task

To use the provided systems model of MarsCity to (1) explore potential architectures and (2) select your preferred design.

Collectively, select your preferred design considering:

- **Desirability** – How well the design meets **your individual goals**.
- **Feasibility** – How well the design supports the **overall StarCity mission**.

Instructions

1. **Individually**, read your *Stakeholder Profile*.

2. **Collectively**, use the *system model* to explore design decisions.
   a. Please select *one team member* to be the formal "simulator".

3. 5 minutes before the end of your collective work, discuss as a group to determine your team's **preferred site design architecture** and record which simulation result number represents it.

Background Information

Decision 1 – Population Allocation: Which population groups reside on what levels?

- Options (per sub-population): (1) Residence 4A, (2) Residence 4B, (3) Residence 4C, (4) Residence 4D, (5) Residence 5.

Decision 2 – Transportation Mode: Which modes of transportation should be supported in the long connections attached to each elevator lobby?

- Options: (1) Walk, (2) Moving Walkway, (3) Scooter.

Decision 3 – Space Function: Where should educational, healthcare and cultural spaces be located on the site?

- Options per level: (1) Educational, (2) Healthcare, (3) Cultural

Performance Metrics

Site Design Evaluation Metric	Definition
Diversity Score [%]	The relative measure of population **diversity within spaces on the site**.
Energy Use [kJ]	The **amount of energy being consumed** by people moving through the site and engaging in activities.
Social Interaction Probability [%]	The measure of the likelihood two people within a space will serendipitously "interact" due to their **proximity to each other**.
Site Utilization [%]	The percentage of spaces within the site with people **actively engaging in a functional space activity** at a given time.

Figure A-3: Team Briefing.

Mars Star City Design & Simulation Workshop
Stakeholder Profile

Team ID		**ParticipantID**		**Are you the team simulator? Yes / No**	

Site Performance Engineer

Individual Role Description

Primary Objective:
- Ensure effective use of the site.

As the Mission Performance Engineer, your role is to *focus on the amount of time the population is engaged in productive activities as opposed to simply moving from one space to another*.

What are the Primary and Secondary occupations? What is cultural? Educational and Healthcare?

Warning: Some variation between simulations, even if they have the same settings, should be expected.

Sub-Populations' *Primary* and *Secondary* Occupations

Sub-Population	Size (people)	Primary Occupation	Secondary Occupation
Farmers	75	**Agricultural**	**Cultural**
Industrial Workers	50	**Industrial**	**Educational**
Life Support Workers	25	**Industrial**	**Healthcare**
Healthcare Workers	25	**Industrial**	**Cultural**
Educators	25	**Educational (1)**	**Public Engagement**

Note: Occupations are color-coded to make comparisons more convenient across individual profiles..

Agricultural Industrial	These site-types **cannot** be changed in the Mars colony simulation.

Your Metric of Interest

- (Maximize) **Site Utilization**
  - **Ideally** over 32%
  - **Undesirable** if it is under 31%

Figure A-4: Stakeholder briefing for Mission Performance Engineer.

Mars Star City Design & Simulation Workshop
Stakeholder Profile

Team ID			ParticipantID			Voice Recorder			Are you the team simulator?	
						Smartwatch			Video Number	

Site Power Budget Engineer
Individual Role Description

Primary Objective:
- Ensure effective utilization of constrained power resources.

As the Mission Power Budget Engineer, your role is to *ensure that the ample but scarce power resources of StarCity are appropriately and effectively utilized.*
Every kJ that is spent in daily activity is a kJ of energy that is not available for supporting the expansion of StarCity or critical activities that happen outside of the village (production of in-situ resources).

Transportation Mode ***Speed Multiplier*** and ***Energy Use***

Mode	Agent Travel Speed (Relative)	System Energy Cost *
Walking	1.00x	0 [kJ / use]
Moving Walkway	1.53x	129.2 [kJ / use]
Scooter	7.34x	396 [kJ / use]

* 1 use == movement down one connection.

Your Metric of Interest
- (Minimize) **Energy Use**
  - **Ideally** under 1,100 [kJ]
  - **Undesirable** if it is over 1,500 [kJ]

Figure A-5: Stakeholder briefing document for Power Budget Engineer.

Mars Star City Design & Simulation Workshop
Stakeholder Profile

Team ID		ParticipantID		Are you the team simulator? Yes / No	

Population Health Engineer

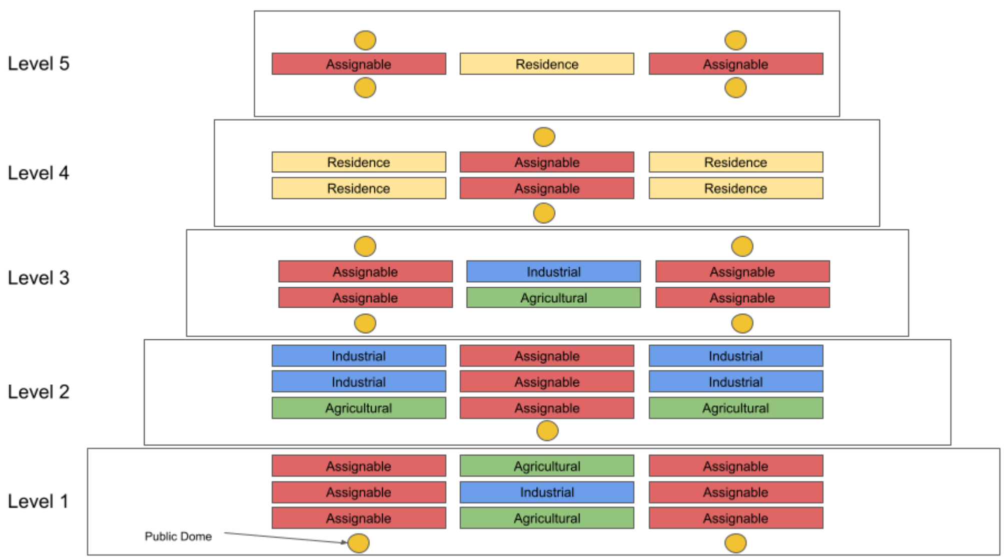
Individual Role Description

Primary Objective: Ensure social interaction of sub-populations within the colony.

As the Mission Health Engineers, your role is to ensure the mental and physical wellbeing of all Martian settlers is maintained. Social engagement is carried out in the public domes. Sub-populations lifestyle level of activity is influenced by the distance among their activities during the day and the method of transportation.
Warning: The Assignable spaces per level are defined all at once. E.g. *Educational space for Level makes all 6 spaces Educational.*

Agricultural Industrial	These site-types **cannot** be changed in the Mars colony simulation.

Top-view breakdown per level



Your Metric of Interest

- (Maximize) **Social Interaction Probability**
  - **Ideally** over 3.5%
  - **Undesirable** if it is under 3%.
- (Maximize) **Diversity Score**
  - **Ideally** over 65%
  - **Undesirable** if it is under 55%.

Figure A-6: Stakeholder briefing for Health Engineer.

Appendix B

Additional Analysis and Tables

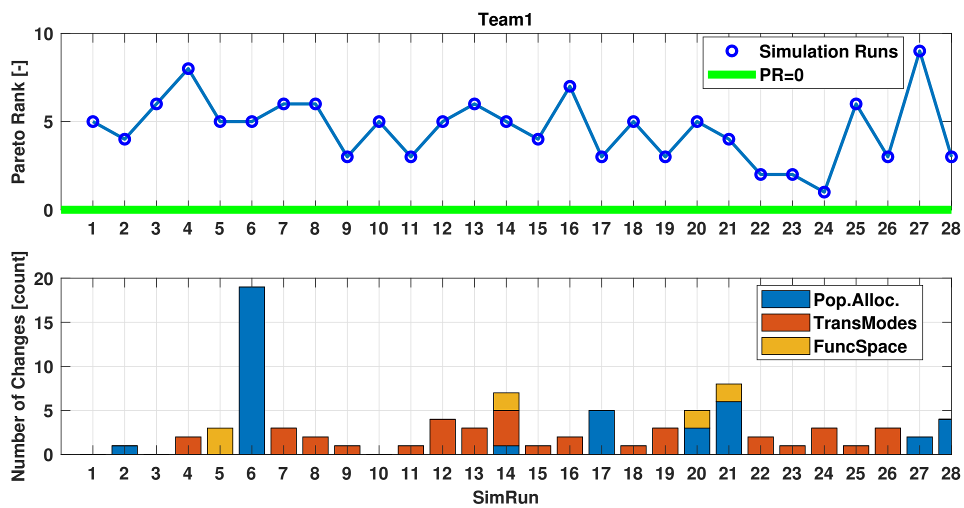


Figure B-1: Pareto Ranking vs Decomposition of Design Changes. This plot shows at the top the Pareto Rank for Team 1, where the x-axis is the number of iterations —in contrast to the *Experiment Time*, as used in most other plots. The green line indicates Pareto Rank = 0, where zero is the most desirable. The bottom plot shows a decomposition of what *types* of changes were made by the team in-between simulation runs. The three colors correspond to the options available in the Mars StarCity user interface. *Note 1*: It is essential to bring to attention that there seem to be moments where the team does a single type –color– of change at the time, and only after having done so a few times does the team start to "mix" the decisions. Whether this is a more general pattern across the teams cannot be deduced from this plot alone.

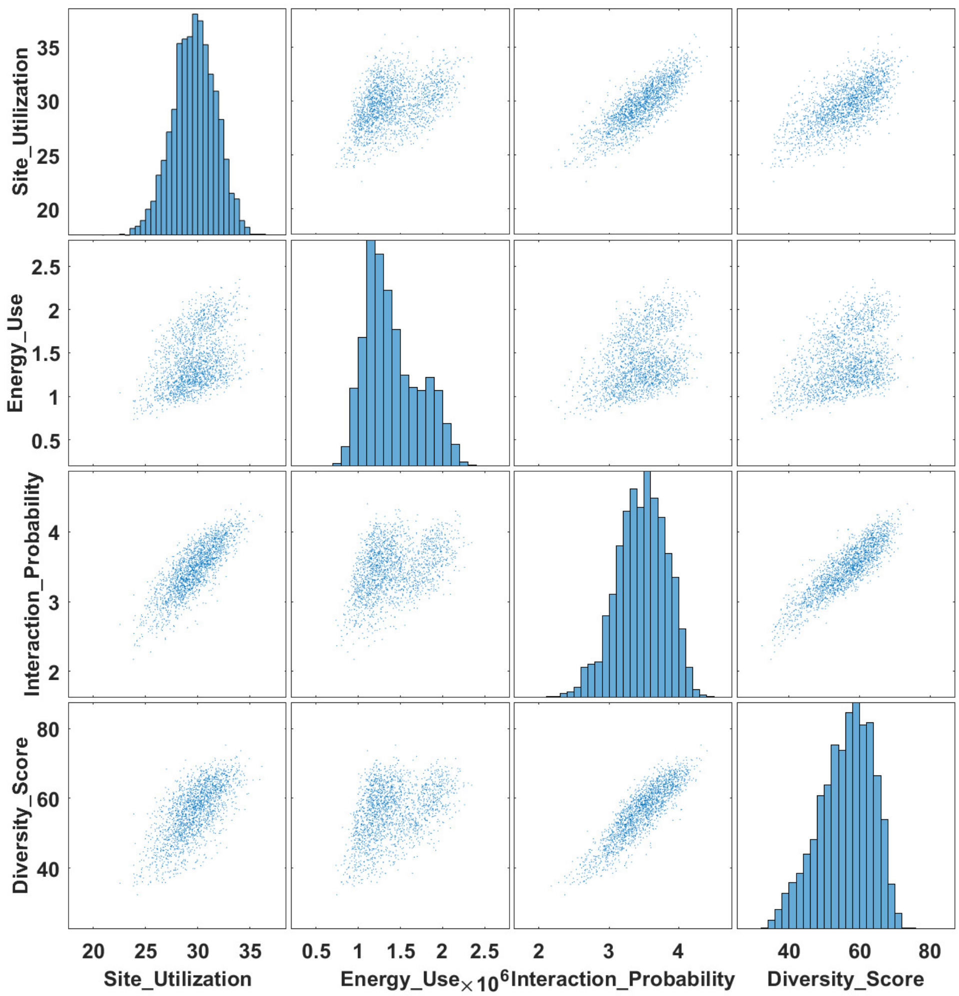


Figure B-2: Correlation Plots for System Metrics.

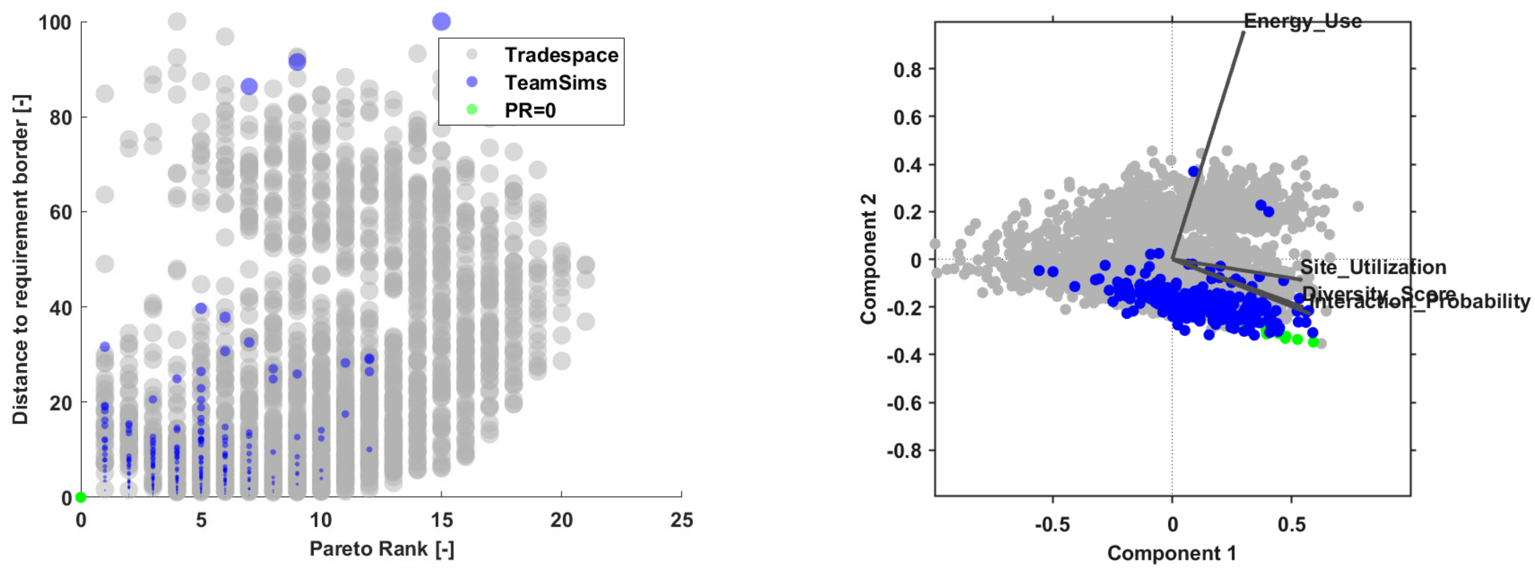


Figure B-3: 4D Tradespace representation.

		Performance			*Emotion*									*ConvPattern*				
Type	**TeamID**	**Ave_PR**	**STD_PR**	**NumSims**	**Ave_ES**	**STD_ES**	**Neutral**	**Fearful**	**Disgusted**	**Sad**	**Surprised**	**Angry**	**Happy**	**dialogue_num sum**	**speech_len sum**	**dia_cv**	**spe_cv**	**speech_len med**
Control	Team01	-1.0	0.5	10	0.030	-0.001	1.8	-0.1	-0.3	-0.2	0.0	-0.9	0.0	-0.43	-0.05	-0.08	-0.06	120
Control	Team03	-0.7	2.6	7	0.049	-0.015	4.9	0.0	0.0	-0.9	0.2	-0.2	1.2	*nan*	*nan*	*nan*	*nan*	*nan*
Control	Team05	-1.4	-0.5	6	0.227	0.037	-4.0	0.1	-0.1	-7.1	1.4	-1.1	2.6	1.06	-0.04	-0.02	-0.16	-245
Control	Team07	-4.1	-5.8	4	0.131	-0.079	3.6	0.1	0.0	-4.2	0.5	-0.8	1.1	-0.56	-0.07	-0.36	-0.26	-1840
Control	Team09	-0.9	1.1	7	0.011	-0.024	6.9	0.0	-0.1	-0.3	-0.7	0.0	0.8	0.76	-0.05	-0.27	-0.23	-1190
Control	Team11	-2.0	-1.7	8	0.021	-0.036	2.2	0.7	-0.1	-1.1	-0.3	0.5	1.3	0.61	-0.05	0.16	0.18	-300
Control	Team13	-0.4	1.2	13	0.098	0.078	6.2	0.2	0.1	-0.1	-0.1	0.5	3.4	2.03	-0.02	0.03	-0.08	180
Treatment	Team02	-4.5	1.3	17	0.098	0.011	3.1	0.6	0.1	-1.3	1.4	-1.4	1.0	3.02	0.03	0.05	-0.05	-975
Treatment	Team04	1.2	2.1	12	0.200	-0.030	13.7	0.1	-0.1	-5.6	0.5	-1.8	-0.1	3.66	0.13	0.13	0.15	-1540
Treatment	Team06	-0.4	-0.9	13	0.120	-0.030	0.7	0.1	0.0	-4.3	0.4	-2.1	-0.7	1.17	0.14	-0.09	-0.06	175
Treatment	Team08	-0.7	0.7	9	0.074	-0.036	5.8	0.1	0.1	0.7	0.2	0.1	1.1	1.36	0.05	-0.03	0.03	430
Treatment	Team10	0.8	-0.2	13	0.026	-0.052	-0.2	-0.1	-0.1	-3.5	-0.5	-0.4	-1.8	0.31	0.10	-0.15	-0.05	1180
Treatment	Team12	2.6	1.3	12	-0.023	0.004	-0.8	0.0	0.1	1.5	-0.1	-1.1	-0.4	1.69	0.09	0.04	-0.01	-680

Figure B-4: Sensemaking factors.

		Physiological Signals			
Type	**TeamID**	**Heart Rate Mean**	**Heart Rate STD**	**Acceleration Mean**	**Acceleration STD**
Control	Team01	0.0	0.0	0.00	0.00
Control	Team03	-0.4	0.0	-120.00	0.00
Control	Team05	1.1	0.0	245.00	0.00
Control	Team07	-0.6	-0.3	1840.00	0.00
Control	Team09	0.8	-0.2	1190.00	0.00
Control	Team11	0.6	0.2	300.00	0.00
Control	Team13	2.0	0.0	-180.00	0.00
Treatment	Team02	3.0	0.0	975.00	0.00
Treatment	Team04	3.7	0.0	1540.00	0.00
Treatment	Team06	1.2	-0.2	-175.00	0.00
Treatment	Team08	1.4	-0.1	-430.00	0.00

Figure B-5: Sensemaking factors - continuation.

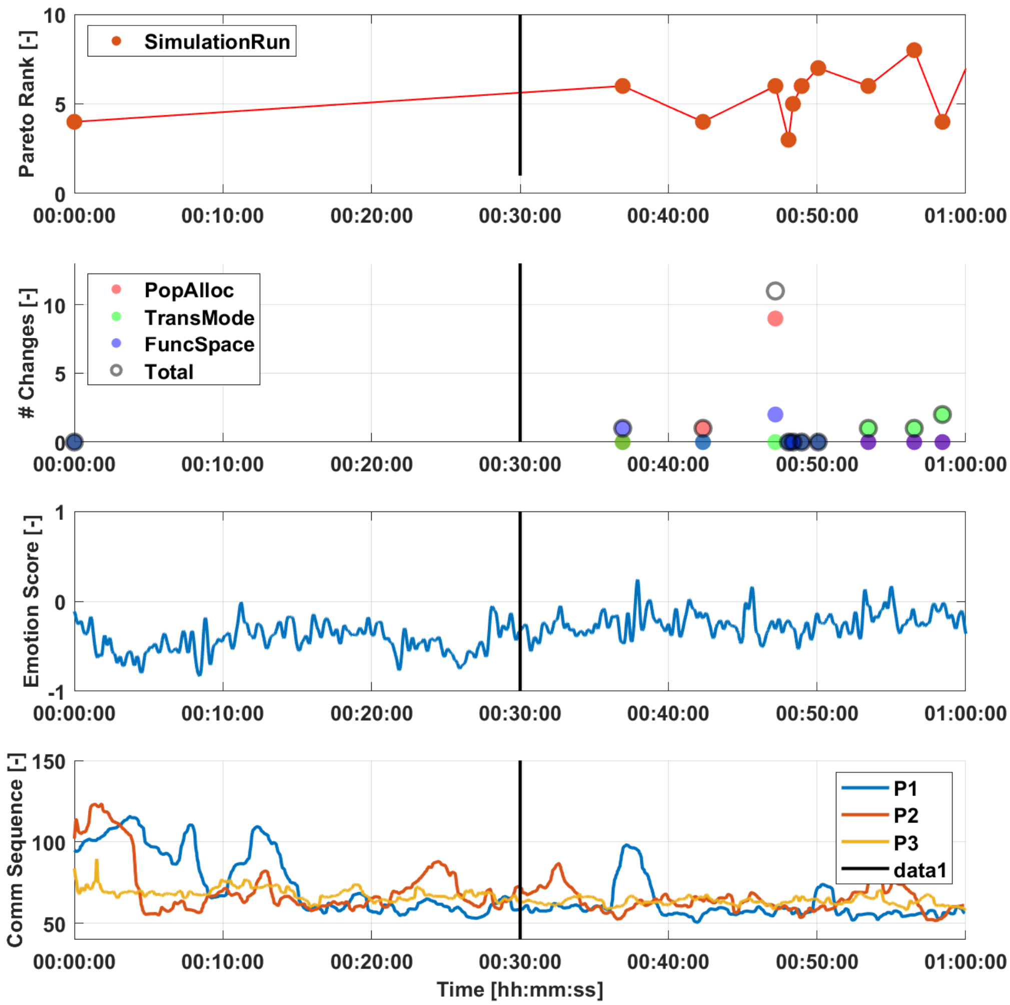


Figure B-6: Multi-modal Team-Level Overview. Each of the subplots shows a different aspect of the team, from the performance at the top, the types and allocation of changes —second from the top, the team's emotion score —third from the top, and the participants' physiological signals —bottom plot. The bold line at the 30 [min] mark denoted the location of the treatment event in the experiment presented in the thesis. *Note 1*: This overview is created from multiple data sources, only one of which —emotion score— can be computed and shown in real-time.

Appendix C

Final Selection Heuristics

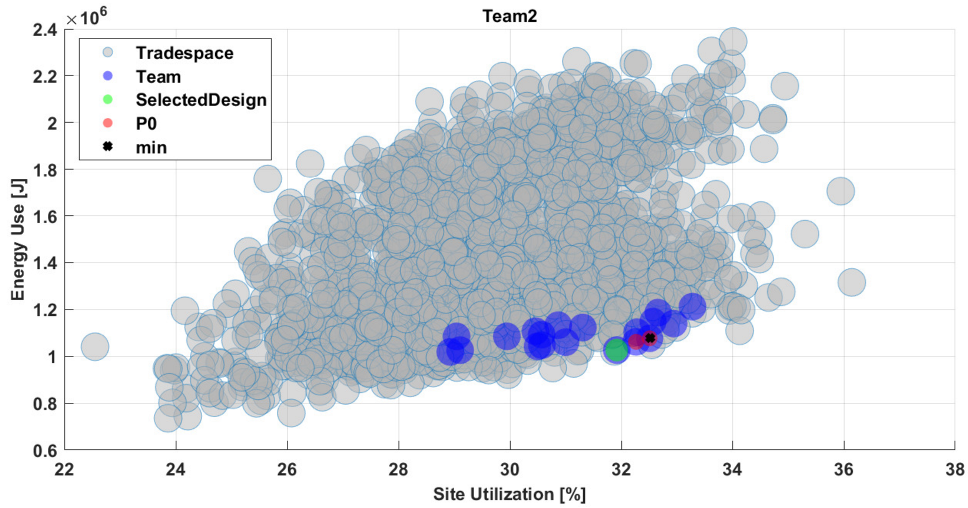


Figure C-1: Team 2 - Selection Heuristic.

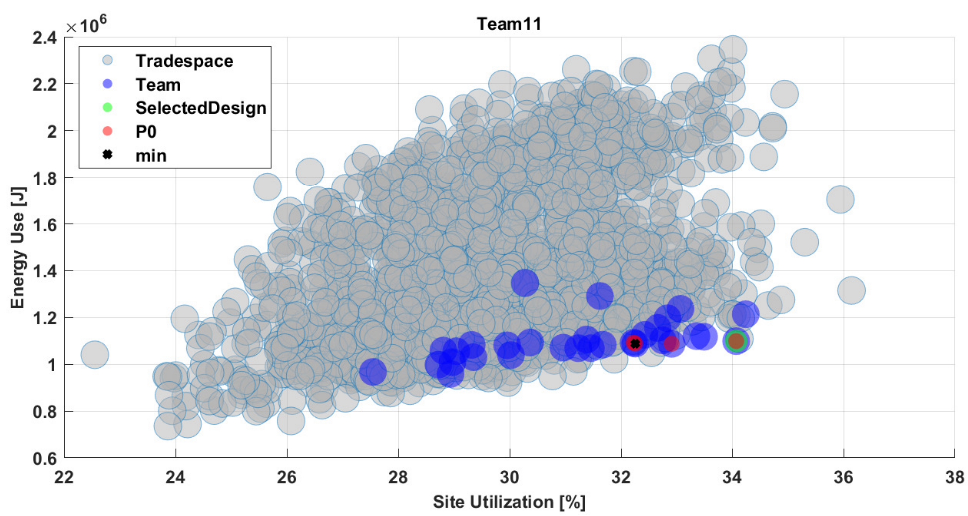


Figure C-2: Team 11 - Selection Heuristic.

1	2	3	4	5	6	7	8
TeamID	**Original Variable Names**	**Selected Design**	**Minimum Distance**	**Delta Selected**	**Delta Minimum**	**SelMetReqs**	**MinMetReqs**
Team1	ParetoFront distance	2.0 45.1	7.0 7.5	2.0 45.1	7.0 7.5	FALSE	FALSE
Team3		0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	TRUE	TRUE
Team5		1.0 30.0	5.0 17.3	1.0 30.0	5.0 17.3	FALSE	FALSE
Team7		2.0 27.7	5.0 9.6	2.0 27.7	5.0 9.6	FALSE	FALSE
Team9		0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	TRUE	TRUE
Team11		0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	TRUE	TRUE
Team13		5.0 26.4	6.0 15.7	5.0 26.4	6.0 15.7	FALSE	FALSE
Team2	ParetoFront distance	1.0 26.4	0.0 0.0	1.0 26.4	0.0 0.0	FALSE	TRUE
Team4		7.0 2.3	7.0 2.3	7.0 2.3	7.0 2.3	FALSE	FALSE
Team6		5.0 11.6	7.0 5.5	5.0 11.6	7.0 5.5	FALSE	FALSE
Team8		1.0 25.0	5.0 10.7	1.0 25.0	5.0 10.7	FALSE	FALSE
Team10		3.0 16.6	6.0 7.2	3.0 16.6	6.0 7.2	FALSE	FALSE
Team12		3.0 31.1	6.0 15.1	3.0 31.1	6.0 15.1	FALSE	FALSE

Figure C-3: Basic Comparison of Selected Designs vs Minimum Distance Designs.

1	2	3	4	5	6	7	8	9	10	11
TeamID	**Original Variable Names**	**Selected Design**	**Minimum Distance**	**Delta Selected**	**Delta Minimum**	**Does Selected Design Meet Goal?**	**Does Minimum Dist Design Meet Goal?**	**Selected Design % Met Goals**	**Minimum Dist % Met Goals**	**Same Selection?**
Team1	Diversity_Score	73.6	62.3	8.6	-2.7	1	0	75.00%	25.00%	FALSE
	Energy_Use	1163327	1200768	63327	100768	0	0			
	Interaction_Probability	4.3	3.8	0.8	0.3	1	1			
	Site_Utilization	34.1	31.2	2.1	-0.8	1	0			
Team3		69.7	69.7	4.7	4.7	1	1	100.00%	100.00%	TRUE
		1085092	1085092	-14908	-14908	1	1			
		4.1	4.1	0.6	0.6	1	1			
		32.3	32.3	0.3	0.3	1	1			
Team5		72.3	62.2	7.3	-2.8	1	0	75.00%	50.00%	FALSE
		1086332	1050504	-13668	-49496	1	1			
		4.1	3.7	0.6	0.2	1	1			
		31.9	29.2	-0.1	-2.8	0	0			
Team7		71.5	64.7	6.5	-0.3	1	0	75.00%	25.00%	FALSE
		1134304	1119856	34304	19856	0	0			
		4.0	3.9	0.5	0.4	1	1			
		32.8	31.2	0.8	-0.8	1	0			
Team9		72.9	70.3	7.9	5.3	1	1	100.00%	100.00%	FALSE
		1088175	1079400	-11825	-20600	1	1			
		4.2	4.2	0.7	0.7	1	1			
		33.3	33.3	1.3	1.3	1	1			
Team11		73.0	70.9	8.0	5.9	1	1	100.00%	100.00%	FALSE
		1098852	1088268	-1148	-11732	1	1			
		4.3	4.0	0.8	0.5	1	1			
		34.1	32.3	2.1	0.3	1	1			
Team13		61.7	57.2	-3.3	-7.8	0	0	50.00%	25.00%	FALSE
		1247235	1160883	147235	60883	0	0			
		4.1	3.6	0.6	0.1	1	1			
		32.3	32.0	0.3	0.0	1	0			
Team2	Diversity_Score	68.9	71.1	3.9	6.1	1	1	75.00%	100.00%	FALSE
	Energy_Use	1026436	1076780	-73564	-23220	1	1			
	Interaction_Probability	4.0	4.3	0.5	0.8	1	1			
	Site_Utilization	31.9	32.5	-0.1	0.5	0	1			
Team4		62.2	62.2	-2.8	-2.8	0	0	25.00%	25.00%	TRUE
		1137648	1137648	37648	37648	0	0			
		3.8	3.8	0.3	0.3	1	1			
		31.0	31.0	-1.0	-1.0	0	0			
Team6		65.9	63.1	0.9	-1.9	1	0	50.00%	25.00%	FALSE
		1118216	1134832	18216	34832	0	0			
		3.6	3.7	0.1	0.2	1	1			
		31.2	30.7	-0.8	-1.3	0	0			
Team8		71.6	66.1	6.6	1.1	1	1	75.00%	50.00%	FALSE
		1039166	1145236	-60834	45236	1	0			
		3.8	3.7	0.3	0.2	1	1			
		29.4	31.6	-2.6	-0.4	0	0			
Team10		66.3	59.9	1.3	-5.1	1	0	75.00%	25.00%	FALSE
		1098196	1154484	-1804	54484	1	0			
		4.0	3.7	0.5	0.2	1	1			
		31.4	31.7	-0.6	-0.3	0	0			
Team12		59.9	55.3	-5.1	-9.7	0	0	50.00%	25.00%	FALSE
		1175706	1100897	75706	897	0	0			
		4.1	3.7	0.6	0.2	1	1			
		34.0	29.9	2.0	-2.1	1	0			

Figure C-4: Detailed Comparison of Selected Designs vs Minimum Distance Designs.

Bibliography

[Abt, 1970] Abt, C. (1987), *Serious Games*. New York: Viking Press.

[Weick, 1979] Weick, K. (1979), *The Social Psychology of Organizing*. New York: McGraw Hill, Inc.

[Hardwig, 1985] Hardwig, J. (1985) *Epistemic Dependence*. The Journal of Philosophy.

[Schwartz & Bilsky, 1987] Schwartz, S., Bilsky, W. (1987), *Toward a Universal Psychological Structure of Human Values*. Journal of Personality and Social Psychology.

[Weick, 1988] Weick, K. (1988), *Enacted sensemaking in crisis situations*. Journal of Management Studies.

[Goldberg, 1989] Goldberg, D. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

[Baker-Brown et al., 1992] Baker-Brown, G., Ballard, E., Bluck, S., De Vries, B., Suedfeld, P., Tetlock, P. (1992), *The conceptual/integrative complexity scoring manual*. Motivation and Personality: Handbook of Thematic Content Analysis (pp. 401-418).

[Levenson, 1992] Levenson, R. (1992) *Autonomic Nervous System Differences Among Emotions*. Psychological Science.

[Russel et al., 1993] Russel, D., Stefik, M., Pirolli, P., Card, S. (1993), *The Cost Structure of Sensemaking*. Conference on Human Factors in Computing Systems.

[Schulman, 1993] Schulman, P. (1993), *The Negotiated Order of Organizational Reliability*. Administration & Society.

[Weick, 1995] Weik, L. (1995), *Sensemaking in Organizations*. Thousand Oaks: Sage Publications.

[Keltner et al., 1997] Keltner, D., Ekman, P. (1997), *Facial Expression of Emotion*. Handbook of Emotions, 2nd edition.

[Brown, 2000] Brown, E. (2000), *Making sense of inquiry sensemaking*. Journal of Management Studies.

[Juarrero, 2000] Juarrero, A. (1999), *Dynamics in Action: Intentional Behavior as a Complex System*. Emergence: Complexity and Organization.

[Sternman, 2000] Sterman, J. (2000), *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Boston: Irwin/McGraw-Hill.

[Wright et al., 2000] Wright, C., Manning, M., Farmer, B., Gilbreath, B. (2000), *Resourceful Sensemaking in Product Development Teams*. Organizational Studies.

[Basu et al., 2001] Basu, S., Choudhury, T., Clarkson, B., Pentland, A. (2001), *Towards Measuring Human Interactions in Conversational Settings*. Proceedings of IEEE CVPR Workshop on Cues in Communication.

[Dolan, 2002] Dolan, R. (2002), *Emotion, Cognition and Behavior*. Science.

[Loewenstein & Lerner, 2003] Loewenstein, G., Lerner, J. (2003), *The Role of Affect in Decision Making*. In R. Davidson, K. Scherer, I. Goldsmith (Eds.), Handbook of Affective Sciences.

[Cooren, 2004] Cooren, F. (2014), *The Communicative Achievement of Collective Minding. Analysis of Board Meeting Excerpts*. Management Communication Quarterly.

[Stein, 2004] Stein, M. (2004), *The Critical Period of Disasters: Insights from Sensemaking and Psychoanalytic Theory*. Human Relations.

[Beal et al., 2005] Beal, D., Weiss, H., Barros, E., MacDermid, S. (2005), *An Episodic Process Model of Affective Influences on Performance*. Journal of Applied Psychology.

[Ilgen et al., 2005] Ilgen, D., Hollenbeck, J., Johnson, M., Jundt, D. (2005), *Teams in Organizations: From Input-Process-Output Models to IMOI Models*. Annu. Rev. Psychol.

[Maitlis, 2005] Maitlis, S. (2005), *The Social Process of Organizational Sensemaking*. Academy of Management Journal.

[Ross & Hastings, 2005] Ross, A., Hastings, D. (2005), *The Tradespace Exploration Paradigm*. INCOSE International Symposium.

[Tetlock, 2005] Tetlock, P. (2005) *Expert Political Judgement: How good is it? How can we know?*. Princeton, N.J: Princeton University Press.

[Weick et al., 2005] Weick, K., Sutcliffe, K., Obstfeld, D. (2005), *Organizing and the Process of Sensemaking*. Organization Science Journal.

[Yang, 2005] Yang, M. (2005), *A Study of Prototypes, Design Activity, and Design Outcome*. Design Studies.

[Blais & Weber, 2006] Blais, A., Weber, E. (2006) *A Domain-Specific Risk-Taking (DOSPERT) Scale for Adult Populations*. Judgement and Decision Making.

[Adler & Obstfeld, 2007] Adler, P., Obstfeld, D. (2007), *The Role of Affect in Creative Projects and Exploratory Search*. Industrial and Corporate Change.

[Ancona et al., 2007] Ancona, D., Malone, T., Orlikowski, W., Senge, P. (2007), *In Praise of the Incomplete Leader*. Harvard Business Review.

[Hindmarsh & Pilnick, 2007] Hindmarsh, J., Pilnick, A. (2007), *Knowing Bodies at Work: Embodiment and Ephemeral Teamwork in Anaesthesia*. Organization Studies Journal.

[Schwartz & Clore, 2007] Schwarz, N., Clore, G. (2007), *Feelings and Phenomenal Experiences*. In E. Higgins & A. Kruglanski (Eds.), Social Psychology: Handbook of Basic Principles.

[Costa & McCrae, 2008] Costa, P., McCrae, R. (2008), *The Revised NEO Personality Inventory*. SAGE Handbook of Personality Theory and Assessment: Volume 2-Personality Measurement and Testing.

[Gloor et al., 2008] Gloor, P., Paasivaara, M., Schoder, D., Willems, P. (2008) *Finding collaborative innovation networks through correlating performance with social network structure*. International Journal of Production Research.

[Grandey, 2008] Grandey, A. (2008), *Emotions at Work: A Review and Research Agenda*. In C Cooper & J. Barling(Eds), Handbook of Organizational Behavior.

[Kim et al., 2008] Kim, T., Chang, A., Holland, L., Pentland, A. (2008), *Meeting Mediator: Enhancing Group Collaborating Using Sociometric Feedback*. Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work.

[Rafaeli et al., 2009] Rafaeli, A., Ravid, S., Cheshin, A. (2009), *Sensemaking in Virtual Teams: The Impact of Emotions and Support Tools on Team Mental Models and Team*. International Review of Industrial and Organizational Psychology.

[Batterman, 2011] Batterman, R. (2011), *The Tyranny of Scales*. The Oxford Handbook of Philosophy of Physics.

[de Weck et al., 2011] de Weck, O., Roos, D., and Magee, C., Vest, C. (2011), *Engineering Systems: Meeting Human Needs in a Complex Technological World*. The MIT Press.

[Pirolli & Russell, 2011] Pirolli, P., Russell, D. (2011), *Introduction to this Special Issue on Sensemaking*. Human-Computer Interaction.

[Ancona, 2012] Ancona, D. (2012), *Sensemaking. Framing and Acting in the Unknown*. The Handbook for Teaching Leadership: Knowing, Doing and Being.

[Gloor et al., 2012] Gloor, P., Grippa, F., Putzke, J., Lassenius, C., Fuehres, H., Fischbach, K., Schoder, D. (2012), *Measuring Social Capital in Creative Teams Through Sociometric Sensors*. International Journal of Organisational Design and Engineering.

[Kim et al., 2012] Kim, T., McFee, E., Olguin, D.O., Waber, B. and Pentland, A. (2012), *Sociometric Badges: Using Sensor Technology to Capture New Forms of Collaboration*. Journal of Organizational Behavior.

[Pentland, 2012] Pentland, A., *The New Science of Building Great Teams*. Harvard Business Review.

[Stigliani & Ravasi, 2012] Stigliani, I., Ravasi, D. (2012) *Organizing Thoughts and Connecting Brains: Material Practices And The Transition From Individual to Group-level Prospective Sensemaking*. Academy of Management Journal.

[Weick, 2012] Weick, K. (2012), *Managing the Unexpected: Complexity as Distributed Sensemaking*. In Making Sense of the Organization Volume 2, K.E. Weick (Ed.).

[Graham et al., 2013] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., Ditto, P. (2013), *Moral Foundations Theory:The Pragmatic Validity of Moral Pluralism*. Advances in Experimental Social Psychology.

[Kaplan & Orlikowski] Kaplan, S., Orlikowski, W. (2013), *Temporal Work in Strategy Making*. Organization Science.

[Maitlis et al., 2013] Maitlis, S., Vogus, T., Lawrence, T. (2013), *Sensemaking and Emotion in Organizations*. Organizational Psychology Review.

[Sosa & Marle, 2013] Sosa, M., Marle, F. (2013), *Assembling Creative Teams in New Product Development Using Creative Team Familiarity*. Journal of Mechanical Design.

[Liu & Maitlis, 2014] Liu, F., Maitlis, S. (2014), *Emotional Dynamics and Strategizing Processes: A study of Strategic Conversations in Top Team Meetings*. Journal of Management Studies.

[Maitlis & Christianson, 2014] Maitlis, S., Christianson, M. (2014), *Sensemaking in Organizations: Taking Stock and Moving Forward*. The Academy of Management Annals.

[Paasivaara et al., 2014] Paasivaara, M., Heikkila, V., Lassenius, C., Toivola, T. (2014), *Teaching Students Scrum Using LEGO Blocks*. In Companion Proveedings of the 36th International Conference on Software Engineering.

[Crawley et al., 2015] Crawley, E., Cameron, B., Selva, D. (2015), *System Architecture: Strategy and Product Development for Complex Systems*. Prentice Hall.

[Lynam & Fletcher, 2015] Lynam, T., Fletcher, C. (2015), *Sensemaking: A Complexity Perspective*. Ecology and Society.

[Vivian et al., 2015] Vivian, R., Tarmazdi, H., Falkner, K., Falkner, N., Szabo, C. (2015), *The Development of a Dashboard Tool for Visualizing Online Teamwork Discussions*. 37th IEEE International Conference on Software Engineering.

[Fruehling & Moser, 2018] Fruehling, C., Moser, B. (2018), *Analyzing Awareness, Decision, and Outcome Sequences of Project Design Groups: A Platform for Instrumentation of Workshop-based Experiments*. International Conference on Complex Systems Design and Management.

[Kozlowski & Chao, 2018] Kozlowski, S., Chao, G. (2018), *Unpacking Team Process Dynamics and Emergent Phenomena: Challenges, Conceptual Advances, and Innovative Methods*. American Psychologist Journal.

[Mathieu et al., 2018] Mathieu, J., Wolfson, M., Park, S. (2018), *The Evolution of Work Team Research since Hawthorne*. American Psychologist.

[Pelegrin, 2018] Pelegrin, L. (2018), *Teamwork Phenomena: Exploring Path Dependency and Learning in Teams During Architectural Design of Sustainable Maritime Shipping Systems*. Master thesis, Massachusetts Institute of Technology, Cambridge, US.

[Pelegrin et al., 2018] Pelegrin, L., Moser, B., Wanaka, S., Chavy-MacDonald, M., Winder, I. (2018), *Field Guide for Interpreting Engineering Team Behavior with Sensor Data*. 2018 Complex Systems Design & Management Conference.

[Rooney et al., 2018] Rooney-Varga, J., Sterman, J., Fracassi, E., Franck, T., Kapmeier, F., Kurker, V., et al. (2018), *Combining Role-play with Interactive Simulation to Motivate Informed Climate Action: Evidence from the World Climate Simulation*. PLoS ONE 13(8).

[Lordos & Lordos, 2019] Lordos, G., Lordos, A. (2019), *Star City: Designing a Settlement on Mars*. 22nd Annual Mars Society Convention.

[Moser et al., 2019] Moser, B., Bellew, C., Cadario, A., Filippi, R. (2019), *The Interplay of Site and Construction Systems Decisions for Complex Urban Infrastructures*. Transdisciplinary Engineering for Complex Socio-Technical Systems.

[Pelegrin et al., 2019] Pelegrin, L., Moser, B., Sakhrani, V. (2019), *Exposing Attention-Decision-Learning Cycles in Engineering Project Teams Through Collaborative Design Experiments*. Proceedings of the 52nd Hawaii International Conference on System Sciences.

[Tan & Moser, 2019] Tan, PS., Moser, B. (2019), *Detection of Teamwork Behavior as Meaningful Exploration of Tradespace During Project Design*. In Cardin, M., et al, (Eds.) Complex Systems Design & Management Asia, Advances in Intelligent Systems and Computing.

[Gaggiolo et al., 2020] Gaggioli, A., Mazzoni, E., Benvenuti, M., Galimberti, C., Bova, A., Brivio, E., Cipresso, P., Riva, G., Chirico, A. (2020), *Networked Flow in Creative Collaboration: A Mixed Method Study*. Creativity Research Journal.

[Gorman et al., 2020] Gorman, J., Grimm, D., Stevens, R., Galloway, T., Willemsen-Dunlap, A., Haplin, D. (2020), *Measuring Real-Time Team Cognition During Team Training*. Human Factors.

[Stefanini et al., 2020] Stefanini, A., Aloini, D., Gloor, P. (2020), *Silence is Golden: The Role of Team Coordination in Health Operations*. International Journal of Operations and Production Management.

[Manandhar, 2020] Manandhar, P. (2020), *Towards Measuring Attention Allocation in Model-based Engineering Teamwork*. Master thesis, Massachusetts Institute of Technology, Cambridge, US.

[Manandhar et al., 2020] Manandhar, P., Rong, K., Carrol, K., de Filippi, R., Winder, I., Dieffenbach, J., Moser, B. (2020), *Sensing Systemic Awareness and Performance of Teams During Model-based Site Design*. 2020 IEEE 6th World Forum on Internet of Things.

[Sanchez et al., 2020] Sanchez-Margallo, J., Gloor, P., Campos, J., Sanchez-Margallo, F. (2020), *Measuring Workload and Performance of Surgeons Using Body Sensors of Smartwatches*. In: Przegalinska, A., Grippa, F., Gloor, P. (Eds.) Digital Transformation of Collaboration. COINs 2019. Springer Proceedings in Complexity.

[Sun et al., 2020] Sun, L., Gloor, P., Stein, M., Eirich, J., Wen, Q. (2020), *"No Pain No Gain": Predicting Creativity Through Body Signals*. In: Przegalinska, A., Grippa, F., Gloor, P. (Eds.) Digital Transformation of Collaboration. COINs 2019. Springer Proceedings in Complexity.

[Roessler & Gloor, 2020] Roessler, J., Gloor, P. (2020), *Measuring Happiness Increases Happiness*. Journal of Computational Social Science.

[Carroll, 2021] Carroll, K. (2021), *Agent-Based Modeling of Population Activity in Complex Terrestrial and Martian Sites*. Master thesis, Massachusetts Institute of Technology, Cambridge, US.

[Gloor et al., 2021] Gloor, P., Fronzetti Colladon, A., Altuntas, E., Cetinkaya, C., Kaiser, M., Ripperger, L., Schaefer, T. (2021), *Your Face Mirrors Your Deepest Beliefs - Predicting Personality and Morals through Facial Emotion Recognition*. Future Internet Journal.

[Lin-Greenberg et al., 2021] Lin-Greenberg, E., Pauly, R., Schneider, J. (2021), *Wargaming for International Relations Research*. European Journal of International Relations.

[Page et al., 2021] Page, P., Karaus, K., Donner, M. (2021),*Enhancing Quality of Virtual Meetings Through Facial and Vocal Emotion Recognition*. [Unpublished Manuscript]

[Riedl et al., 2021] Riedl C., Kim Y., Gupta P., Malone T., Woolley A. (2021), *Quantifying Collective Intelligence in Human Groups*. Proceedings of the National Academy of Sciences of the United States of America.

[McDonough, 2021] McDonough, K. (2021), *Detecting the Influence of Stakeholders' Mental Models on Emergent Collective Awareness in Instrumented Teamwork Workshops*. Master thesis, Massachusetts Institute of Technology, Cambridge, US.

[Wiltshire at al., 2021] Wiltshire, T., Hudson, D., Lijdsman, P., Wever, S., Atzmueller, M. (2021) *Social Analytics of Team Interaction Using Dynamic Complexity Heat Maps and Network Visualizations*. arXiv:2009.04445v2.

[Wolbers, 2021] Wolbers, J. (2021), *Understanding Distributed Sensemaking in Crisis Management: The Case of the Utrecht Terrorist Attack*. Journal of Contingencies and Crisis Management.

[Gloor et al., 2022] Gloor, P., Zylka, M., Fronzetti Colladon, A., Makai, M. (2022), *Entanglement - A New Dynamic Metric to Measure Team Flow*. Social Networks, 70, 100–111. [Accepted manuscript.]

[Krehbiel, 2022] Krehbiel, N. (2022), *Stakeholder mental model alignment influence on mid-stage performance of new product engineering teams*. Master thesis, Massachusetts Institute of Technology, Cambridge, US.

[Peng et al., 2022a] Peng, S., Amakasu, T., Horri, H. (2022a), *Teamwork Visualization for Team Reflection Based on Changes in Communication Patterns*. Journal of Industrial Integration and Management.

[Peng et al., 2022b] Peng, S., Komatsuzaki, S. (2022b), *Temporal Equal and Active Participation in Synchronous Collaborative Learning: Antecedents and Effect for Learning*. [Unpublished Manuscript].

[Peng et al., 2022c] Peng, S. (In Press) *Physiological Synchrony in Virtual Teams: Prediction of Team Emergent States*. 82nd Annual Meeting of the Academy of Management.

[Vazquez et al., 2022] Vazquez, I., Faruque, F., Moser, B. (2022), *Instrumenting Weick's Seven Properties to Measure Collective Sensemaking in Engineering Teams*. The 29th International Society of Transdisciplinary Engineering.